Old Dominion University

ODU Digital Commons

Engineering Management & Systems Engineering Theses & Dissertations

Engineering Management & Systems Engineering

Fall 2004

Application of Queuing Theory and Procedure Time Estimation in a Local Healthcare System

Galina Tsoy
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/emse_etds

 Part of the Operational Research Commons, Quality Improvement Commons, Systems Engineering Commons, and the Systems Science Commons

Recommended Citation

Tsoy, Galina. "Application of Queuing Theory and Procedure Time Estimation in a Local Healthcare System" (2004). Master of Science (MS), Thesis, Engineering Management & Systems Engineering, Old Dominion University, DOI: 10.25777/vv57-7857
https://digitalcommons.odu.edu/emse_etds/201

This Thesis is brought to you for free and open access by the Engineering Management & Systems Engineering at ODU Digital Commons. It has been accepted for inclusion in Engineering Management & Systems Engineering Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

APPLICATION OF QUEUING THEORY AND PROCEDURE TIME ESTIMATION IN A LOCAL HEALTHCARE SYSTEM

by

Galina Tsoy

B.S. December 2002, Old Dominion University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

MASTER OF SCIENCE

ENGINEERING MANAGEMENT AND SYSTEMS ENGINEERING

OLD DOMINION UNIVERSITY

December 2004

Approved by:

DR.RABADI (Director)

DR.UNAL (Member)

DR.MUN (Member)

ABSTRACT

APPLICATION OF QUEUING THEORY AND PROCEDURE TIME ESTIMATION IN A LOCAL HEALTHCARE SYSTEM

Galina Tsoy
Old Dominion University, 2004
Director: Dr.Ghaith Rabadi

People in the United States pay more for their basic healthcare needs than do people in any other nation in the world. When we consider that the United States is the wealthiest nation in the world, controlling the majority of the world's resources, it seems only reasonable to ask: Why should it be this way?

In an effort to address this problem, this thesis examines two possible methods of improving health care efficiency in hospitals. The thesis is thus in two parts: the first part examines resource allocation in medical units using Queuing Theory, and the second part examines a more accurate estimation of surgical procedure times. In the first part, a queuing model provides performance measures based on the historical interarrival and service times of a medical unit. The queuing model demonstrates the trade-off between the utilization (system's perspective) and patient waiting time in the queue (customer's perspective). Also, it shows some insights as to the average of patients spent in the system and in the queue, the average time a patient spends in the system, and the probability of the system being empty. The queuing model will enable hospital managers to see the effect of arrival rate, service rate, and number of beds to estimate the main

performance measures of assessing the benefits of providing extra beds to minimize patient waiting time when demand increases.

The second part provides a better estimate of surgical procedure times based on a lognormal distribution. Efficient estimation of surgical procedures times will reduce the costs incurred in inaccurate estimation of its time and consequently the costs associated with surgical operating rooms.

These two proposed methodological approaches will hopefully point the way toward further research aimed at bringing about concrete improvements in U.S. hospital performance.

©, 2004, Galina Tsoy, All Rights Reserved.

It is often said that when "the student is ready the teacher appears." This thesis is dedicated to Dr. Ghaith Rabadi, whose guidance, support, and encouragement has shown me that the reverse may also be true, and that appearances may precede readiness. Thank you for readying me to move forward.

ACKNOWLEDGMENTS

The completion of this thesis represents one milestone in a long journey that would have been unimaginable without the help and generosity of some special people whom I am privileged now to acknowledge: First and above all, I want to thank my parents, Natalya and George Tsoy for granting me the independence to travel so far from home and the confidence to live out my dreams.

I want also to thank Dr. Ghaith Rabadi for believing in me and for showing me what it means to maintain one's scientific vision and focus. I only hope that this thesis measures up, in some estimable way, to his high standards. In addition, I am indebted to my committee members, Dr. Resit Unal, who has been a source of unwavering encouragement since I arrived at Old Dominion University five years ago, and Dr. Ji Hyon Mun, whom I have come to regard as an outstanding role model and mentor. At Sentara Leigh Hospital, Terry McKenna and Julie Sigler reliably and courteously provided me with the raw data that made this entire project feasible. In addition, I want to thank Samuel Kovacic whose advice and confidence in my abilities has been a source of significant strength to me.

Writing the text to accompany this project, in a language that is neither native nor entirely familiar to me, has been a true challenge, one that has given me a deep appreciation of the richness of the English language, as well as a deep admiration of those for whom this wealth is second nature. I am grateful to my partner, Dana Heller, for inviting me into this resplendent world, and for remaining by my side every step of the way.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

I. INTRODUCTION

Hospitals are the single largest source of healthcare costs in the United States, and despite widespread concern there appears to be no end in sight to rising costs. With an aging population of "baby-boomers" and the nation's growing number of uninsured, the system is facing a crisis of serious proportion. Calls for government regulation along with growing market competition have put increasing pressure on hospitals to achieve better performance at the lowest possible cost. Patients, as they continue to evolve into customers, are requesting shorter waits in hospitals, quicker turnaround of results, and better quality of care. For years, hospitals have been exploring ways to improve their margins and their patients' satisfaction. As other service industries do, healthcare organizations shall satisfy patient demands for better service while simultaneously controlling the costs of resources. However, despite these efforts patients still wait long hours in hospitals to receive very costly service.

A big part of the waiting problem is supply of adequate number of hospital beds. Green and Nguyen (2001) showed that patient waiting time is strongly related to inadequate bed resources. To allocate bed resources, hospital managers generally employ a simple approach to determine bed capacity, an approach which is based on target occupancy level. Hospital bed occupancy is defined as the ratio of occupied beds

to the total number of beds. The most frequently employed target level is 85 % (Green, 2002).

The original goal of target occupancy level was to control the number of hospital beds to minimize the cost. The target occupancy level was developed in 1970s at the federal government level as a response to elevating hospital costs, and was based on estimates of "acceptable" delays (McClure 1976). There are several problems with reported occupancy level. First, internal data for calculations typically include certified beds and beds "in service" while reported occupancy is based on only certified or licensed beds. Second, published occupancy levels typically are based on the average "midnight census" which measures the lowest occupancy level of the day. Finally, reported occupancy levels are yearly averages, and therefore do not reflect the seasonality and variability between weekdays and weekends. For all these problems, published occupancy levels are not reliable measures for determining the bed utilization (Green, 2002). Therefore, the approach of calculating bed resources based on target occupancy level leads to the false results and thus to rising cost of healthcare service.

To decrease the rising cost of healthcare, health managers call for a reduction in bed capacity. However, by doing this, managers create long patient waiting times. From the system perspective, performance of a system is measured by the system throughput and resource utilization (bed utilization). From the customer perspective, performance of a system is based on the server response time (minimum waiting time). To satisfy these conflicting goals, managers need to consider many factors, including costs, the probability of turning patients away, waits for emergency patients, backups for transferred patients from other departments, and patient dissatisfaction. The insight on

some of the mentioned factors can be provided by queuing theory. Queuing models provide useful information for designing and evaluating the performance of queuing systems. They involve trade-offs between server utilization and waiting time/delays, which the target occupancy level lacks to provide.

Many mathematical methods and stochastic approaches have been studied. There are two main categories: 1. Analytic models which include a stochastic process that describes the flow of patients and utilization of resources, 2. Simulation models which include hypothetical and empirical data to imitate the hospital system. With respect to the healthcare industry, a wide variety of analytic and simulated queuing models (Gross and Harris, 1998) are available to help healthcare managers evaluate queuing theory. Some of the early work was presented by Bailey (1952) and Welch (1964) in modeling appointment systems in outpatient facilities. Gupta et al (1971) used the basic multiple-server queuing model for staffing hospital care units in inpatient settings. Bennett and Worthington (1998) discussed the organizational challenges found in the implementation of queuing models in an outpatient clinic.

Some queuing model attempts have been made to suit a variety of different types of care units, such as obstetric services, operating rooms, trauma centers, cardiac care units, geriatric units, and emergency departments. In each case, queuing theory is used to determine the appropriate allocation of resources. For example, Milliken et al. (1972) used a queuing model to predict room utilization. Taylor (1969) and Tucker et al. (1999) used queuing theory in operating rooms. Dexter and Macario (2001) determined the optimal number of beds and occupancy of a unit to minimize staffing costs in obstetrics units. Green and Nguyen (2001) examined data to estimate bed availability in intensive

care units and obstetrics. Litvak et al. (2004) validated a queuing model in busy intensive care units (ICU).

The growing body of using queuing theory is evidence of the increasing recognition of the relevance and value of Queuing Theory in addressing the problems currently faced by the healthcare industry. However, compared with many other organizations, hospitals are still reluctant and have been slow in adopting queuing theory as a means to improve their performance. Even when proposed queuing models are relevant and reliable, the results are not always used and applications are scattered. Therefore, in order to test and observe the applicability of queuing theory in hospitals, this thesis will focus on validation and credibility of a queuing model in the medical units of Sentara Leigh Hospital, in Norfolk, Virginia. Customarily, patients who have been diagnosed with cardiac disease are referred to these medical units. The results from the queuing model will show some insights as to bed capacity and bed utilization, waiting time, the average of patients for the unit being studied.

CHAPTER II

II. LITERATURE REVIEW

This literature review consists of three parts. Part 1 "The history of queuing theory" primarily based on *Fundamentals of Queuing Theory* (Gross and Harris, 1998), *Queuing Methods* (Hall, 1991), and *Management Science* (Taylor III, 2002). Part 2 describes queuing theory applications in healthcare industry. Part 3 provides a summary of the chapter.

HISTORY OF QUEUING HISTORY

The history of queues goes back to primitive life; an early queue was described in the Bible. Despite the fact that queue existed for centuries, queuing theory is quite modern. Only in the beginning of twentieth century queuing theory was developed by Danish mathematician and statistician Agner Krarup Erlang who determined the number of switches and minimum waiting time to place a call. He published in 1909 his pioneering paper "The theory of probabilities and telephone conversations" on the study of congestion of telephone traffics. Particularly notable was Erlang's argument for the input of telephone calls to obey the Poisson law. Using this rationale in his most significant paper, "Solution of some problems in theory of probabilities of significance in automatic telephone exchanges", Erlang also assumed here that service times are exponential. Much of his work that followed during the period to the Second World War in 1940 was related to the design of automatic telephone exchanges. Among other

contributions, Erlang studied queuing in the M/M/1/∞ system and loss model M/M/m/0 applying "birth-and-death" representations. After a few years, British Post Office established Erlang's formula as the basis for calculations circuit facilities.

Although the first use of the term "queuing theory" did not occur until 1951, in an article by David G. Kendall that appeared *in The Journal of the Royal Statistics*, there were many early pioneers of queuing theory whose work has been summarized (Saaty 1961). Some of those pioneers are O'Dell (1920), Fry (1928), Molina (1927), Kolmogorov (1931), Khinchin (1932), and Crommelin (1932). Inspired by Erlang's work on queuing, Pollaczek (in the 1930s through the 1960s) developed the formula for a single channel with Poisson arrivals and arbitrary service time which is called the Pollaczek-Khintchine formula. He also studied the constant service time for ordered queue discipline and for allocation by subqueues in front of each server. Pollaczek developed a formula the general service time with general arrivals and multiple channels (G/G/c).

The considerable growth of Queuing Studies appeared with the finding of Operation Research in the late 1940s and early 1950s. The first textbook was written by Morse in 1958 called "Queues, Inventories, and Maintenance". In 1961, another book was written by Saaty "Elements of Queuing Theory with Applications, and in 1976, "Queuing Systems" was completed by Kleinrock. To day more than 100 books have been published on Queuing Theory.

Over the last 40 years, research on Queuing Theory has progressed significantly. Journals such as *Operation Research, Naval Research Logistics Quarterly, European Journal of Operational Research and Management Science* often include contributions

on this subject. Now *Journals of Applied Probability, Advances in Applied Probability, and Queueing Systems: Theory and Applications, Stochastic Models, and Probability in the Engineering and Information Sciences* are specifically oriented journals to queuing theory.

There are many real life applications of queuing theory in industries such as telecommunications, banking, copy business, airlines, and police (Brigandi et al. 1994; Brusco, Jacobs, Bongiorno, et al. 1995; Brigham 1955, and Taylor and Huxley 1989). Unlike optimization theory where the objective function needs to be minimized or maximized subject to constraints, queuing theory is mostly a mathematical descriptive theory. It formulates, interprets and predicts performance measures to better understand a queuing system. It provides an insight into the cost of a system and customers delays, waits, and their satisfaction in terms of response time.

QUEUING THEORY IN THE HEALTHCARE INDUSTRY

Weiss and McClain (1986) described the process by which patients are discharged from acute care units. Sometimes patients wait in acute units for extended services and occupy the beds not for medical reasons. The collected data of seven hospitals in New-York State was used to validate the proposed queuing model. An M/G/∞ model where M denotes Poisson arrivals, G denotes a general service time, and ∞ is infinite number of servers was applied. The problem with this model is that it is only applicable to one acute facility and one extended facility. The advantage is simplicity of the model and that the decision maker may have a greater control over the decision variables for one hospital rather for all facilities in a region. Based on this model, census distributions can be

estimated which will predict the average of patients from acute care and the maximum expected impact of patients on census predictions for the entire hospital. Then hospital administrators can evaluate a cost/benefit using quality-of-care measures. Also, the distributions of the time patients spent waiting for extended care can be predicted.

Worthington (1987) applied an M/G/S model. In this model, arrivals occur at random at a rate λ_q, service rate is general and independent from any probability distribution, and S is number of servers. The model assumed that arrivals are random at a rate that decreases linearly with waiting list size. The author considered eight examples: the present system, a waiting list without feedback, increasing the number of beds, decreasing mean service time, combining two equal lists, combining two unequally resourced lists, introducing feedback earlier, and finally, combining two differently managed lists. Statistical evidence suggested that observed process during an 18-month period is adequately described by the model.

Siddharthan et al. (1996) investigated the increasing patient waiting time costs and proposed an economic solution to deal with this problem at a public and non-profit hospital. The data was provided by a large acute facility in Dade County, Florida for three months in the fall 1989. The patients were classified by two classes: emergency and non-emergency care. Non-emergency patients who can get a service in other facilities than in emergencies impose a burden to a hospital by creating a longer waiting line and time. Emergency patients are prioritized over non-emergency patients and prioritizing includes preemption or suspending a service for non-emergency patients. Among non-emergency patients, the service was on a first come, first in basis. The authors have

proved that using a priority queuing model reduces the waiting times for both emergency and planned patients in emergency care.

El-Darzi et al. (1998) proposed a simulation and flow model to assess the effect of blockage, occupancy, and emptiness on patient flow in geriatric inpatient department. They considered the model as a queuing system where interarrival rate was exponential (the arrivals of the next patient was independent from the previous), the service time followed exponential distribution, and the queuing discipline was First in First out (FIFO). The patients arrive at the acute compartment if beds are available, if not, they are rejected. Queue 1 is created when no beds are available in rehabilitation compartment for the patients arriving from acute stay. Those patients occupy beds in the acute compartment. Queue 2 is created similarly when patients from rehabilitation compartment are needed to be transferred to long-stay compartments. The number of beds for each compartment was determined from the simulation results. The results from the flow and the simulation model proved that they were viable tools to determine bed occupancy in a geriatric department.

Tucker et al. (1999) used queuing theory to determine whether a trauma center needs an additional OR room during the night time period. The objective was to provide necessary service while minimizing patient waiting time. The single-phase, single-server (M/M/1) basic formula from queuing theory was applied. The probability of two or more patients needing the operating room at the same time will reflect the possibility of needing a back up night shift. Then, simulation was developed and its results were compared to the queuing model to validate the results from the queuing model. The data included one-year operating room cases. It turned out that the probability of two or more

cases needing the OR at the same time is 0.1%; therefore trauma operating rooms at nighttime did not need a backup team.

Kim et al. (1999) provided a study of queuing theory and simulation models with actual data from intensive care unit (ICU) facility in Hong Kong over six-month period. The ICU received patients from four different sources: ward; accidents and emergency; Operation Theatre (OT) –emergency; and Operation Theater (OT) – elective. There was only one queue for all different sources. The arrival rate and service rate was proven to be Poisson and exponential except for elective surgery. The Chi-square goodness-of-fit did not support the exponential assumptions for OT - electives because patients usually come from scheduled appointments. Even though for OT - electives, the service time did not follow the exponential distribution, it was still assumed that the service rate is exponential; however, in the calculations, the actual variance rate calculated from the data was used. A simulation model was used to validate the results of the queuing theory. Together, the analytical (queuing) model and simulation provided the same results and some insights into managerial aspects of ICU's operations.

Gorunescu et al. (2002) integrated queuing theory with compartmental models of flow to show how arrivals, length of stay, and number of beds influence bed occupancy, emptiness and turn away in geriatric departments in a London teaching hospital of geriatric medicine. Also the authors showed how availability of unstaffed beds influences the turn way and cost. An M/PH/c/N queuing model was used, where M denotes Poisson arrivals, PH denotes phase-type service distribution, and c is the number of beds, and N is the maximum capacity of the system. The service distribution is considered as a mixture where the components of the mixture are characterized by the phase of discharge (acute,

rehabilitative, long stay) from the system. Five scenarios were tested for the queuing model: 1. changing the arrival rate, 2. changing length of stay, 3. changing the bed allocation, 4. adding a five or ten bedded waiting rooms, 5. costing the policy. These modeled scenarios showed the effect of arrivals, service rates, number of beds, and extra beds on probability of rejection. Using the queueing model, the authors provided the methodology that enables decision makers to estimate main characteristics of access to inpatient beds and assess the benefits of providing extra beds to minimize turn away when demand increases. Further work is needed to evaluate this methodology in a real life situation.

Green (2002) used queuing theory to estimate bed unavailability in ICU and obstetrics units. The analysis was based on 1997 data for obstetrics and intensive care units for New-York state hospitals and it illustrated how the target occupancy levels may be misleading and potentially dangerous. She argued hospital occupancy needs to be determined by queuing rather than by target occupancy level. A standard M/M/s queuing model (Gross and Harris 1985) was utilized to determine different performance measures, specifically the waiting time/delay in queue for service to show the disadvantage of using target occupancy level. The delay from acute and obstetric units was measured from the time a bed was requested to the time at which a bed was available. Then it was compared with probability of delay from the queuing model. The queuing model provided better results than target occupancy level. Also, the queuing model gave more factors affecting the trade off between utilization and the number of beds. From the model, for a given occupancy level, delays increase as number of beds decreases.

Litvak (2004) sought to validate the utility of queuing theory in an intensive care unit (ICU). They examined an 18-bed unit of an urban children hospital in the medical-surgical ICU during a 2-year period. An M/M/c/s model was applied where the arrival time is Poisson, service time is exponential, and there is c number of servers and s number of spaces in the system. The queuing discipline is first-come, first-served (FCFS). The observed monthly arrival rates, available beds, and stay time, monthly utilizations and rejection probabilities were determined applying a formula for M/M/c/s and compared with observed turn-away rates and utilizations. The results suggested that queuing theory provides very accurate information to predict turn-away rates and represents a simple and reasonable approach.

SUMMARY

Based on the literature review of the work that has been done in obstetrics units, operating rooms, trauma centers, geriatric, and intensive care units, queuing theory is shown to be valid approach for estimating bed capacity, bed occupancy, and waiting time. This thesis uses queuing theory in medical units of Sentara Leigh Hospital, in Norfolk, Virginia. Customarily, patients who have been diagnosed with cardiac disease are referred to these medical units. The queuing model will provide the insights to see the trade-off between utilization (system perspective) and patient waiting time (customer perspective). An M/M/c model where the arrival time is Poisson, service time is exponential and independent from other arrivals, and there is c number of beds will be applied here.

CHAPTER III

III. QUEUING MODEL

Different types of Queuing models:

The simplest form of queuing system is the single server with a single waiting line or queue. A store can have one server (e.g. cashier) and consequently one waiting line or queue. Another form of queuing models is the multiple-server model where several independent servers in parallel serve a single waiting line. These two types are the most common in queuing systems.

Among these two general categories (Hall, 1991), there are models that consider the following aspects:

- Balking: when customers estimate how long they may have to wait for a service and leave immediately
- Reneging: when customers do not join the line but leave later
- Jockeying: when customers move between the lines.

Balking, reneging, and jockeying are the most difficult aspects of a queuing system to measure because it is hard to record customers in the system: will a customer return back or never, if he/she returns and when will it be. In this thesis, these aspects are not captured since they do not directly apply here.

QUEUING CHARACTERISTICS

A queuing system must have the following characteristics:

1. Customer. A customer is a person who waits for a service: shoppers, patients, bank customers, etc. The arrival process of customers can be in person or in groups, at a constant rate or in a pattern, predictable or random. In addition, different customers may arrive at different times of the day.
2. Servers. A server is a person or resource of serving a customer: doctors, bank tellers, beds, etc. The service time of a server can be constant or varied; dependent on the type of customer; predicted in advance; and dependent on the server or the time of the day. In addition, service time is dependent whether identical or different tasks are performed; and what the rules are for moving customers from server to another.
3. Queuing Discipline. The queuing discipline specifies the order in which customers are served. Customers can be served on a first-in, first-out (FIFO) basis (most common), or on a last –come, first-served (LCFS) basis, or a last-in, first – out (LIFO). Also, queuing can be random when parts/customers are selected randomly, or prearranged when customers arrive according to prearranged scheduled appointments, or processed alphabetically according to customers' last names, such as school registration or at job interviews. Other disciplines exist including those with priorities.

MEASURES OF PERFORMANCE

Customer Measures of Performance

Customer measures of performance in most cases are waiting time in the queue and cost associated with this waiting time. Evidently, a customer wants to spend less time in the queue and in the system. There is a cost associated with the waiting time, which could be for some customers more costly than others. In addition to quantitative measures, qualitative measures are important to the customers such as the waiting environment, whether the customers are informed about their waiting time, whether they can sit, and whether the room is crowded.

Server Measures of Performance

The server measure of performance is the cost of providing service, which is reflected in utilization and service time. A short service time is an indication of more efficiency. The utilization of a server is the percentage of time the server is busy. From a server perspective, a longer queue length is more costly since more space is required to accommodate more customers.

KENDALL NOTATION

Kendall notation named after the statistician Kendall (1953) is a shorthand notation to identification and description of the systems.

Kendall notation has the form *A/B/c/K/m/Z* where

A is the interarrival time distribution,

B is the service time distribution,

c is the number of servers,

K is the largest possible number of customers in the queue (i.e. queue size)

m is the number of customers in the source,

Z is the method by which the queue is serviced (i.e. queue discipline).

In most common real-world systems, this notation can be shortened to *A/B/c,* where *K* and *m* are assumed to be infinite, and Z is assumed to be first-in, first-out. Common symbols representing time distributions for *A* and *B* include D for deterministic (or constant) time distribution; M (Markovian) for exponential time distribution, which corresponds with random arrivals; G for general service time distribution; GI for general independent interarrival time distribution, and Ek for k-Erlang.

Examples of queuing systems include

- M/M/c denotes a multiple-server queue with an exponential interarrival distribution (M), with exponential service time (M), and c servers
- M/G/1 denotes a single server queue with general service time distribution (G) and exponential interarrivals (M)
- D/G/c denotes a multiple server queue with deterministic (constant) interarrival and general service rate.

DEFINITIONS

Here, healthcare terminology will be used.

Discharge time is the time the patient leaves the system.

Departure time from queue is the time the patient leaves the queue to be served.

Time in queue is the departure time from queue minus the arrival time.

Time in system is the discharge time minus the arrival time or time in queue plus service time.

The arrival process

This subsection consists of the description of arrival rate definition, Poisson arrival pattern and interarrival times.

Arrival Rate Definition

An arrival at unit *i* is recorded when a patient is referred to unit *i* by a doctor. The arrival rate (λ) is the frequency at which customers (patients) arrive at a waiting line according to a probability distribution. This rate can be estimated from historical data derived from studying the system. Although arrivals can be described by any probability distribution, the arrival rate is often defined by a Poisson distribution.

Poisson arrival distribution

Here are some important characteristics of the Poisson distribution:

1. Poisson distribution is a discrete distribution where the random variable is limited to a set of distinct non-negative values.
2. Expected value (E) and variance (V) are the same and equal λt (t is the time units of the next arrival).
3. The probability of the next arrival or probability of no arrival during the next t units of time is equal $e^{-\lambda t}$.

Interarrival times of Poisson process

When the arrival times follow a Poisson distribution, it is known that the interarrival time follows the exponential distribution (Hall, 1991).

Characteristics of the exponential distribution:

- The exponential distribution is continuous and defined over the set of non-negative real numbers.
- The expected value (E) or mean is equal to $1/\lambda$
- The variance (V) is equal $1/\lambda^2$.

The service time

The service time (μ) denotes the length of time that a patient physically spends in a system while being served. The recording of the service time thus starts at the time of physical transfer to unit i. The service rate is the average number of patients who can be served during a time period. Like arrival times, service times need to be defined by a probability distribution. Although service times can be described by any probability distribution, they are often defined by an exponential distribution. The mean service time per bed can be shown as an exponential expression with its mean $1/\mu$ and the variance $1/\mu^2$. The service and the arrival times must have compatible units of measurement.

DATA COLLECTION

Data was collected over the thirty three days by Sentara Leigh Hospital. The historical data was available for at least sixth month period but it did not capture the time we were interested in. For example, the queuing time in many cases was combined with service times. For that purpose, the data collected was the maximum days we could obtain from Sentara Leigh Hospital. Over the thirty three days, 307 points was collected which was reasonable for validation of the queuing model.

M/M/c MODEL

Our analysis is based on an M/M/c queuing model to estimate different performance measures. Three important elements for the queuing model must be determined: the average admissions/day arrival rate (λ), the service rate (μ), and number of beds (c). The Queuing discipline is First-In, First-Out (FIFO). Based on the historical data of Sentara Leigh Hospital over thirty-three days, the interarrival rate is occurred according exponential distribution and the service rate has an exponential process as well. The number of beds (servers) can be varied to determine the trade-off between utilization and patient waiting time.

Performance measures of a system:

The probability that there are no customers in the system (Taylor III, 2002) is

$$Po = \frac{1}{\left[\sum_{n=0}^{n=c-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c\left(\frac{c\mu}{c\mu-\lambda}\right)} \qquad 3.1$$

The average number of customers in the system is

$$L = \frac{\lambda\mu(\lambda/\mu)^c}{(c-1)!(c\mu-\lambda)^2} Po + \frac{\lambda}{\mu} \quad 3.2$$

The average time a customer spends in the system is

$$W = \frac{L}{\lambda} \quad 3.3$$

The average number of customers in the queue is

$$L_q = L - \frac{\lambda}{\mu}$$

The average time a customer spends in the queue waiting to be served is

$$W_q = W - \frac{1}{\mu} = \frac{Lq}{\lambda} \quad 3.4$$

$$\text{The utilization} = \frac{\lambda}{c\mu} \quad 3.5$$

Our estimated arrival rate, service rate, and number of beds are based on Sentara's historical data. Every 24 hours, on the average 9.3 patients arrive to the hospital. The average service time is 87.36 hours.

Arrival rate (λ)	9.3/24 = 0.38	patient/hour
Service rate ($1/\mu$)	1/87.36 = 0.0115	1/hour
Number of servers (c)	41	beds

Substituting these values in equations 3.1 to 3.5, we get the following:

The probability of no patients in the system is

$$Po = \frac{1}{\left[\sum_{n=0}^{41=41-1} \frac{1}{41!}\left(\frac{0.3875}{0.0115}\right)^{41}\right] + \frac{1}{41!}\left(\frac{0.3875}{0.0115}\right)^{c}\left(\frac{41*0.0115}{41*0.0115-0.3875\lambda}\right)} = 0.00\ \%$$

The average number of customers in the system is

$$L = \frac{0.3875*0.0115(0.3875/0.0115)^{41}}{(41-1)!(41*0.0115-0.3875)^2}\ 0.00 + \frac{0.3875}{0.0115} = 34.65$$

The average time a customer spends in the system is

$$W = \frac{L}{\lambda} = 34.65 / 0.3875 = 89.41$$

The average number of customers in the queue is

$$L_q = L - \frac{\lambda}{\mu} = 34.65 - \frac{0.3875}{0.0115} = 0.8$$

The average time a customer spends in the queue waiting to be served is

$$W_q = W - \frac{1}{\mu} = \frac{Lq}{\lambda} = 0.795 / 0.3875 = 2.05$$

$$\text{The utilization} = \frac{\lambda}{c\mu} = \frac{0.3875}{41*0.0115} = 0.83 \text{ or } 83\ \%.$$

A summary is provided in a Table 1.

Table 1 Performance measures of the queuing model.

Average server utilization	83	%
Average number of customers in the queue (L_q)	0.8	patient
Average number of customers in the system (L)	34.65	patient
Average waiting time in the queue (W_q)	2:03	hour/patient
Average time in the system (W)	89:24	hour/patient

CHAPTER IV

IV. RESULTS AND DISCUSSION

During the recorded thirty three day period, 307 patients were admitted to the medical units of Sentara Leigh Hospital in Norfolk, Virginia. Before using an M/Mc queuing model, the interarrival times and service times were extracted from the historical data to see if they occurred according to an exponential distribution. The distribution fitting was preformed by Arena (simulation software). The graphical representation of the interarrival and service times:

Figure 1 Histogram of interarrival times.

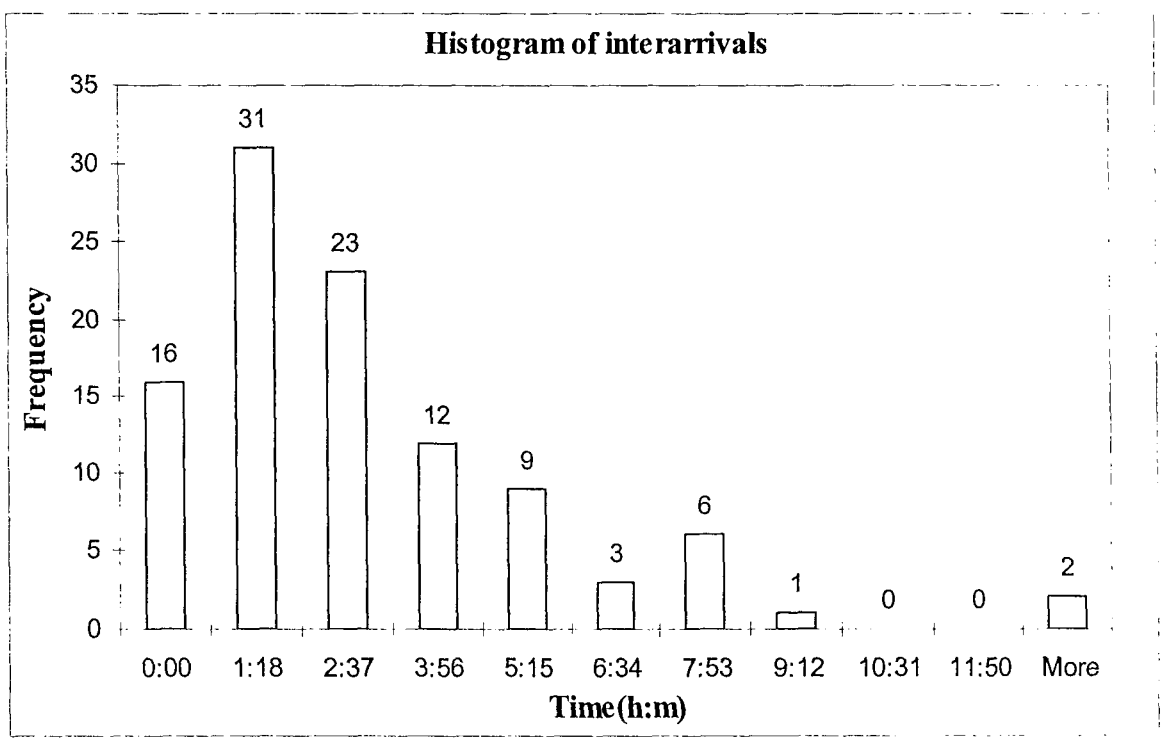
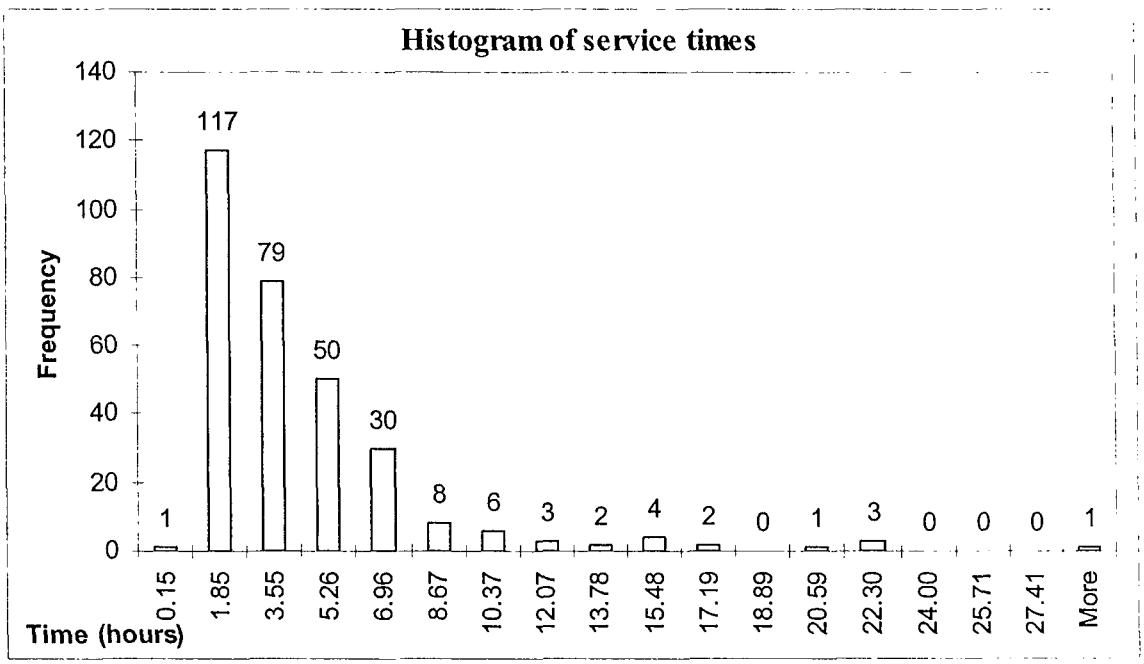


Figure 2 Histogram of service times.



For distribution fittings, two tests were conducted: Kolmogorov-Smirnov (K-S) and Chi-Square. K-S test tries to determine if two datasets differ significantly. The chi-square test is used to test if a sample of data came from a population with a specific distribution. The K-S and Chi-Square tests for both interarrival times and service times did not reject the null hypothesis with the confidence level of 95 percent. The hull hypothesis (Ho) states that the distribution 1 is equal to the distribution 2. The alternative hypothesis (H_1) states that distribution 1 is different from the distribution 2. A p-value of interarrivals for the K-S test was more than 0.15, and for the Chi-square was 0.664. A p-value of service times for he K-S test was 0.01, for Chi-square was 0.05. In both cases, the p value was large and therefore we could not reject Ho. Thus, we conclude that an exponential distribution is a good fit for the data.

After performing the tests for interarrivals and service times, a queuing model validation was conducted.

Queuing model validation

The queuing model was validated by comparing its results with the Sentara Leigh Hospital's s historical data. This was important to be done before conducting any experiments or analysis on the queuing model. The actual bed utilization, the actual average waiting time in the queue and in the system were compared to the model's results.

Table 2 shows a summary of the performance measures based on the historical data.

Table 2 Performance measures of actual (historical data).

Average server utilization	85	%
Average waiting time in the queue (W_q)	2:24	hour/patient
Average time in the system (W)	89:45	hour/patient

Comparing these with the model's output in Table 1, the difference is small and we can conclude that the queuing model is valid.

Relationship between the bed utilization and patient waiting time

To show the relationship between the utilization and patient waiting time, three scenarios were tested for our queuing model: 1. changing the arrival rate, 2. changing service time, 3. changing the number of beds.

Results from the three scenarios:

1. Changing the arrival rate from 7 to 11 patients/24h. Since the actual arrival rate is 9.3 patients/24h, we performed a sensitivity analysis varying arrival rate from 7 to 11 patients/24h. Table 3 and 4 show the performance measures only for those values. The performance measures for the other arrivals are in Appendix A. When the arrival time increases, all performance measures increase. The utilization increases linearly (Fig.5), while the average number of patients and average time spent in the system and in the queue increase exponentially at some point (Fig.3, 4). From the Fig. 3 and 4, it can be seen that the arrival rate 10 patients/24h is the critical point for the system. When the arrival time is 11 patients/24h, the average number of patients and the average time spent in the system and in the queue increases exponentially.

Table 3 Performance measures when arrival rate is 11 patients/24h.

Average server utilization	97.7	%
Average number of customers in the queue(L_q)	34.6	patients
Average number of customers in the system(L)	74.64	patients
Average waiting time in the queue(W_q)	75:29	hour/patient
Average time in the system(W)	162:51	hour/patient

Table 4 Performance measures when arrival rate is 7 patients/24h.

Average server utilization	0.62	%
Average number of customers in the queue(Lq)	0.01	patients
Average number of customers in the system(L)	25.49	patients
Average waiting time in the queue(Wq)	0:01	hour/patient
Average time in the system(W)	87:22	hour/patient

Figure 3 Average number of patients in the system (L) and in the queue (Lq) when arrival time increases.

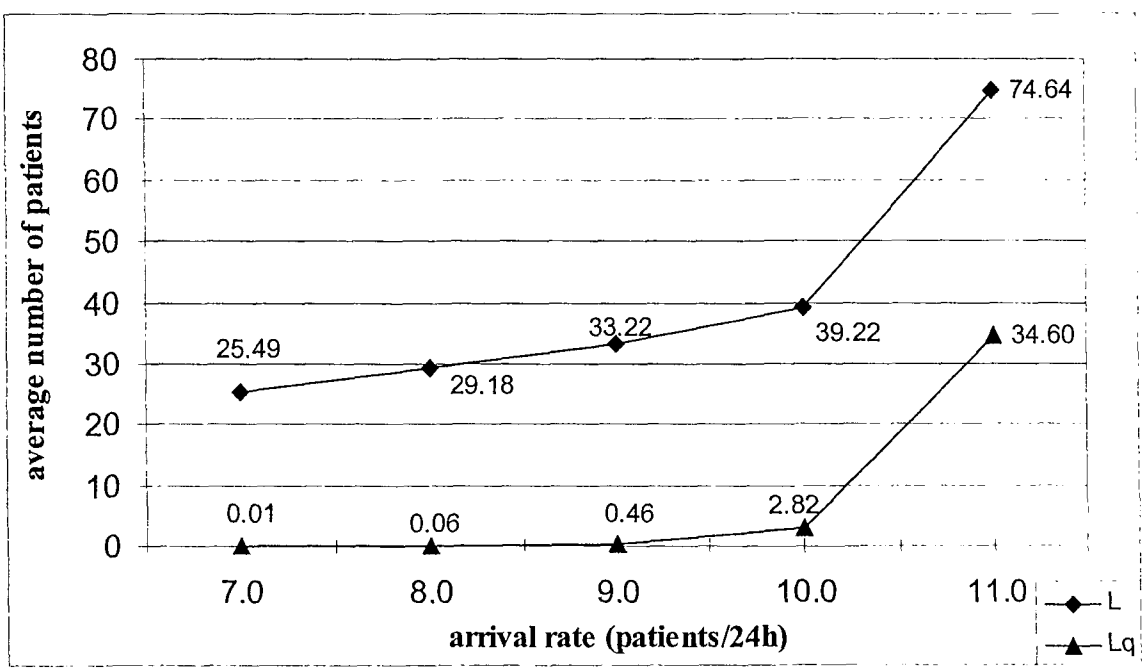


Figure 4 Average time in the system (W) and in the queue (Wq) when number of patients increases.

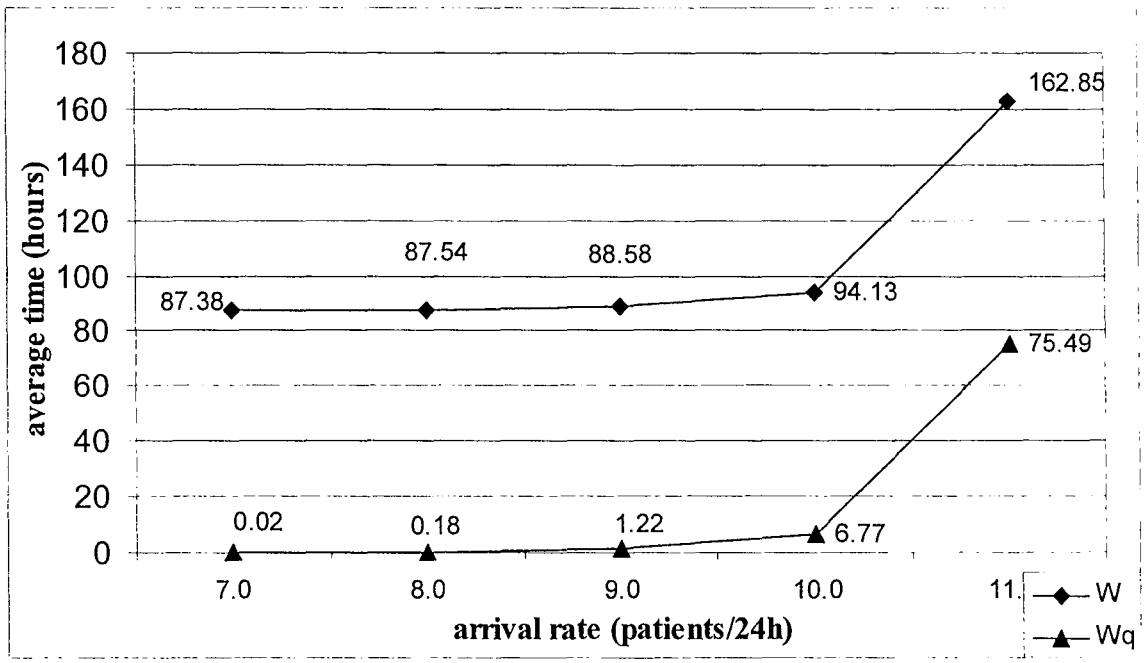
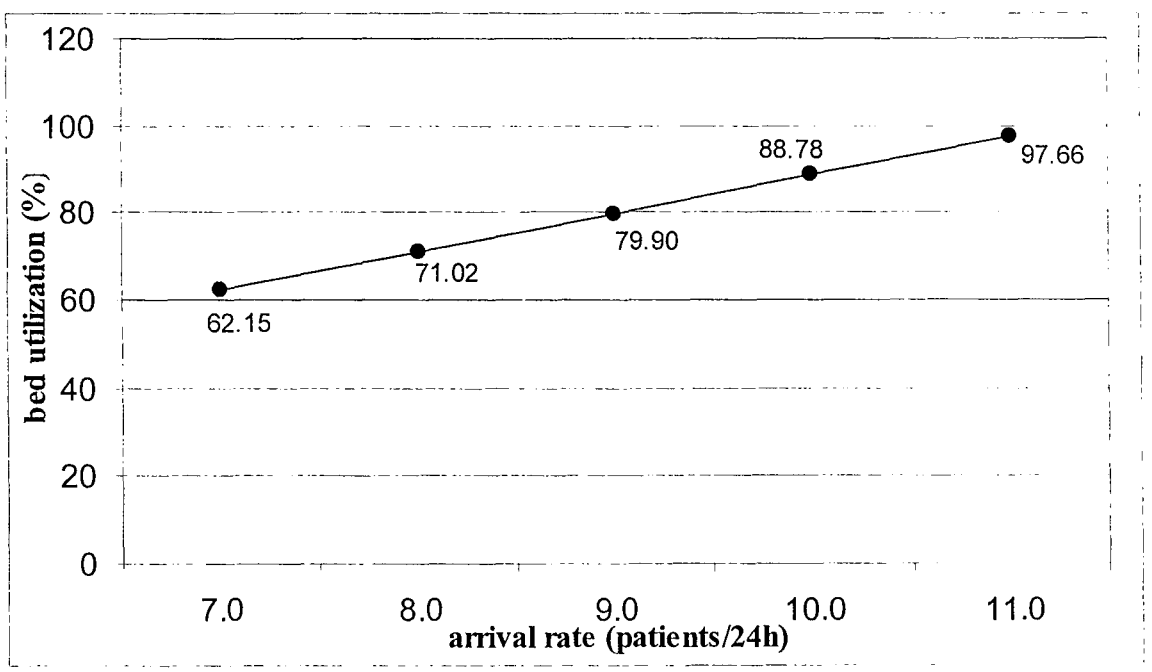


Figure 5 Utilization when number of patients increases.



2. Changing the service time from 2.5 days to 4.3 days/patient. The actual service time is 3.64 days/patient. Table 5 and 6 show the performance measures only for those values. The performance measures for the other days are in Appendix A. When the service time increases, all performance measures increase. The utilization increases linearly (Fig.8), while the average number of patients and average time spent in the system and in the queue increase exponentially at some point (Fig 6, 7). From Fig. 6 and 7, it can be seen that the service rate 4 days/patient is the critical point for the system. When the service time is 4.3 days/patient, the average number of patients and the average time spent in the system and in the queue increases exponentially.

Table 5 Performance measures when service time is 4.3 days/patient.

Average server utilization	97.54	%
Average number of customers in the queue (L_q)	32.52	patients
Average number of customers in the system (L)	72.51	patients
Average waiting time in the queue (W_q)	83:54	hour/patient
Average time in the system (W)	187:06	hour/patient

Table 6 Performance measures when service time is 2.5 days/patient.

Average server utilization	56.71	%
Average number of customers in the queue (L_q)	0.00	patients
Average number of customers in the system (L)	23.25	patients
Average waiting time in the queue (W_q)	0:00	hour/patient
Average time in the system (W) – in hours	60:00	hour/patient

Figure 6 Average number of patients in the system and queue when service time increases.

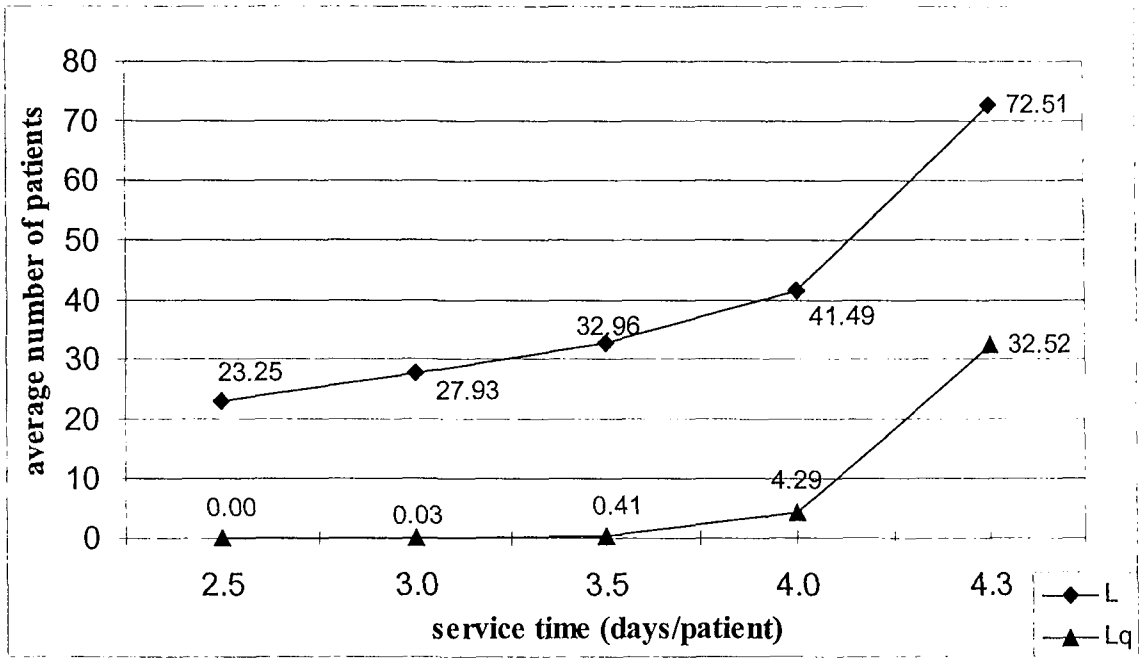


Figure 7 Average number of patients in the system and queue when service time increases.

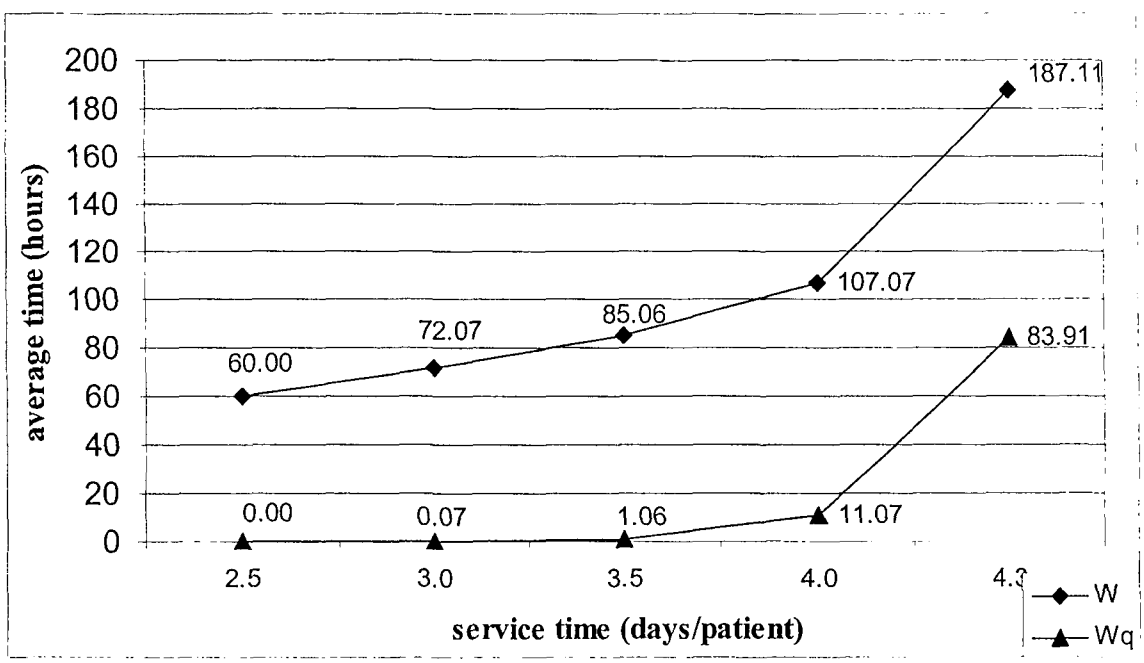
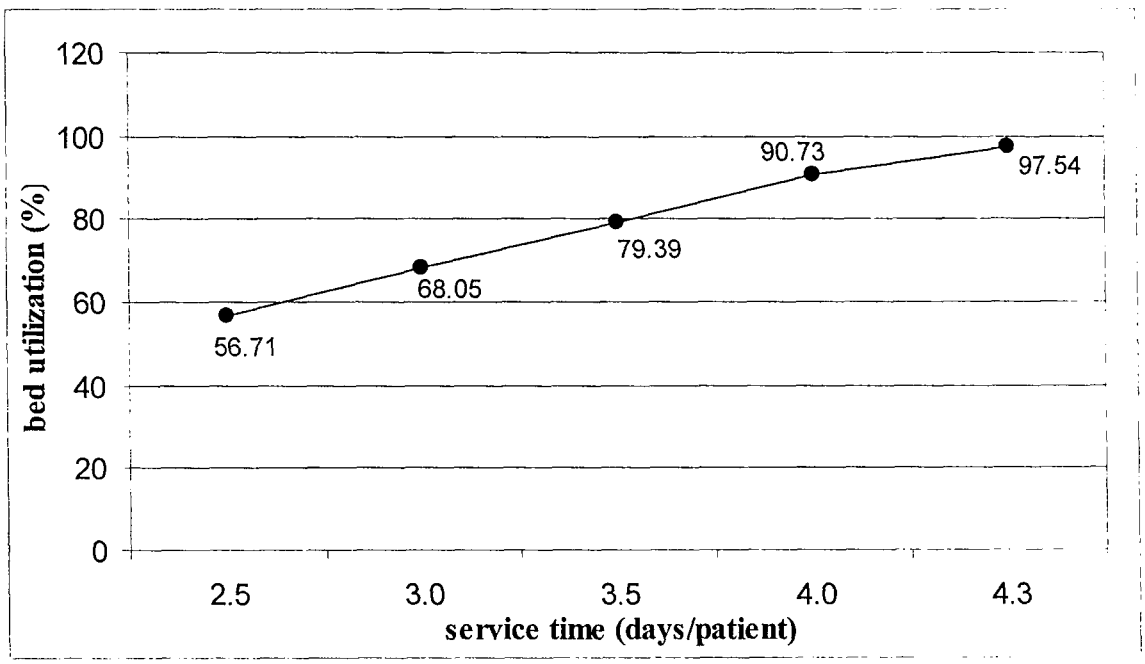


Figure 8 Utilization when service time increases.



3. Changing the number of beds from 34 to 48 beds. In a table 8, the highlighted numbers are the values of the performance measures of the queuing model which correspond to the actual number of beds in the medical units of Sentara Leigh Hospital. When the number of beds decreases, the utilization increases (Fig.9). The average number of patients in the queue and in the system is increasing exponentially. The average waiting time in the queue and in the system is increasing exponentially (Fig.10, 11). Here, it can bee seen that there is a trade-off between the utilization and the patient waiting time. By increasing the number of beds, the utilization decreases linearly while the waiting time increases exponentially. The critical point here is 35 beds (Fig.10, 11).

Table 7 Performance measures when the number of beds decreases.

Number of beds	Utilization	Average patients in Queue (L)	Average patients in system (L)	Average waiting time in Q (Wq)	Average waiting time in system (W) in hours	W (in days)
48	70.53	0.0	33.9	0:05	87:26	3.64
47	72.03	0.1	33.9	0:08	87:30	3.65
46	73.59	0.1	33.9	0:13	87.35	3.65
45	75.23	0.1	34.0	0:21	87:42	3.65
44	76.94	0.2	34.1	0:33	87:54	3.66
43	78.73	0.3	34.2	0:51	88:13	3.68
42	80.60	0.5	34.4	1:19	88:41	3.70
41	**83**	**0.8**	**34.6**	**2:03**	**89:24**	**3.73**
40	84.63	1.2	35.1	3:11	90:33	3.77
39	86.80	1.9	35.8	5:01	92:23	3.85
38	89.08	3.1	37.0	8:06	95:28	3.98
37	91.49	5.3	39.2	13:42	101:04	4.21
36	94.03	9.9	43.7	25:27	112:49	4.70
35	96.72	23.1	57.0	59:37	146:59	6.12
34	99.56	221.8	255.7	572:28	659:49	27.49

Figure 9 Utilization when number of beds decreases.

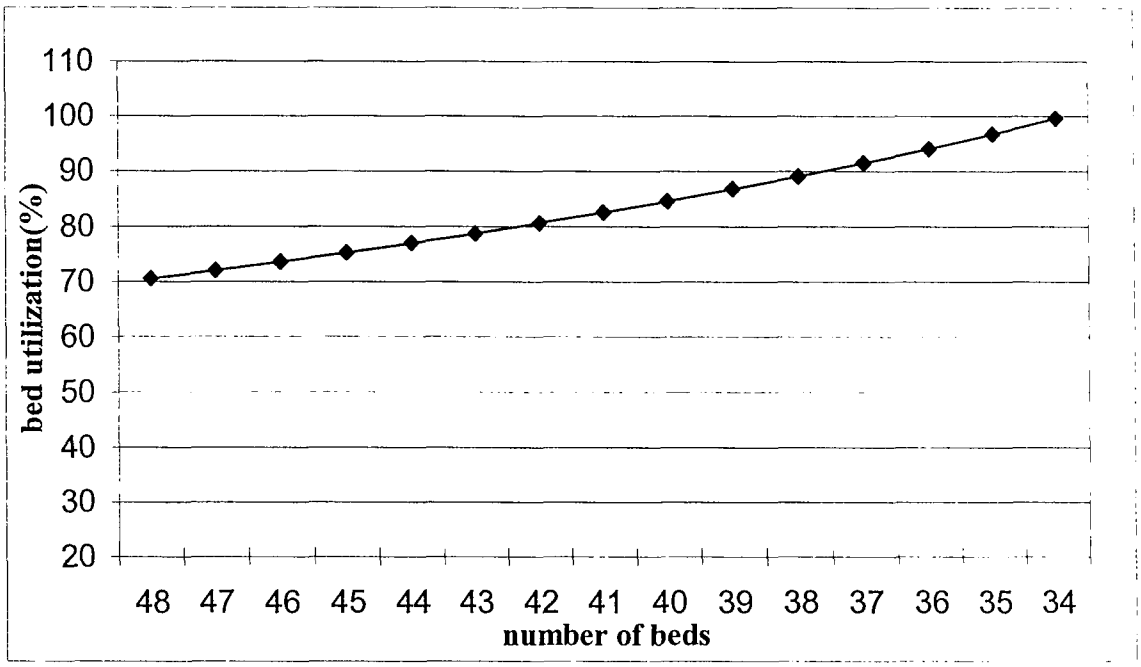


Figure 10 Average number of patients in the system (L) and in the queue (Lq) when number of beds increases.

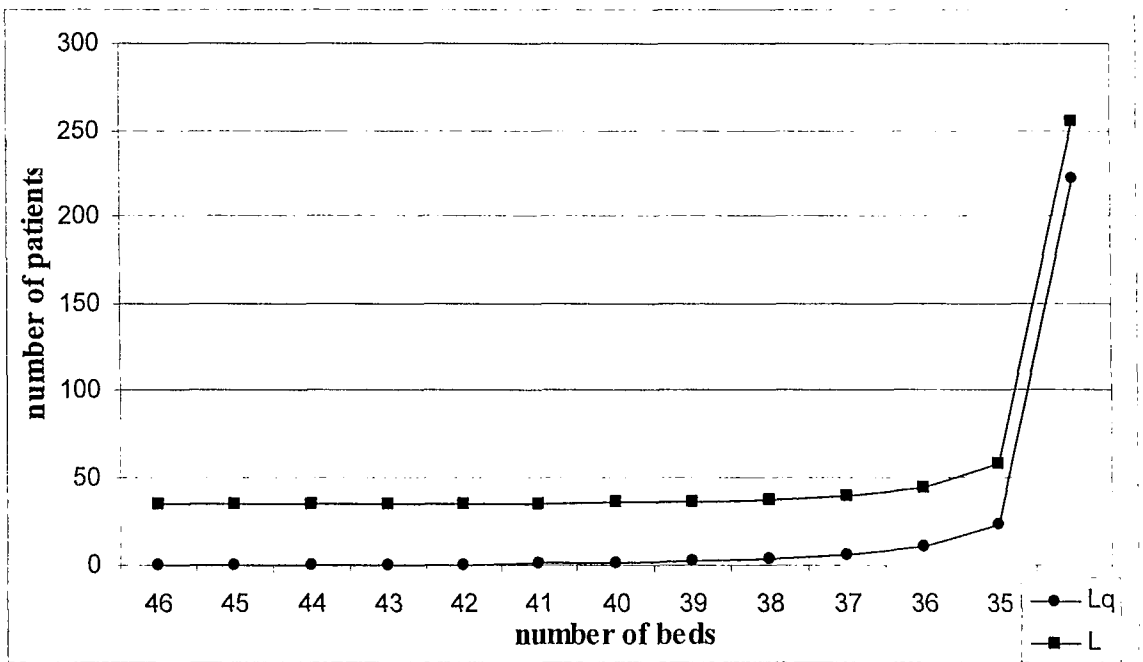
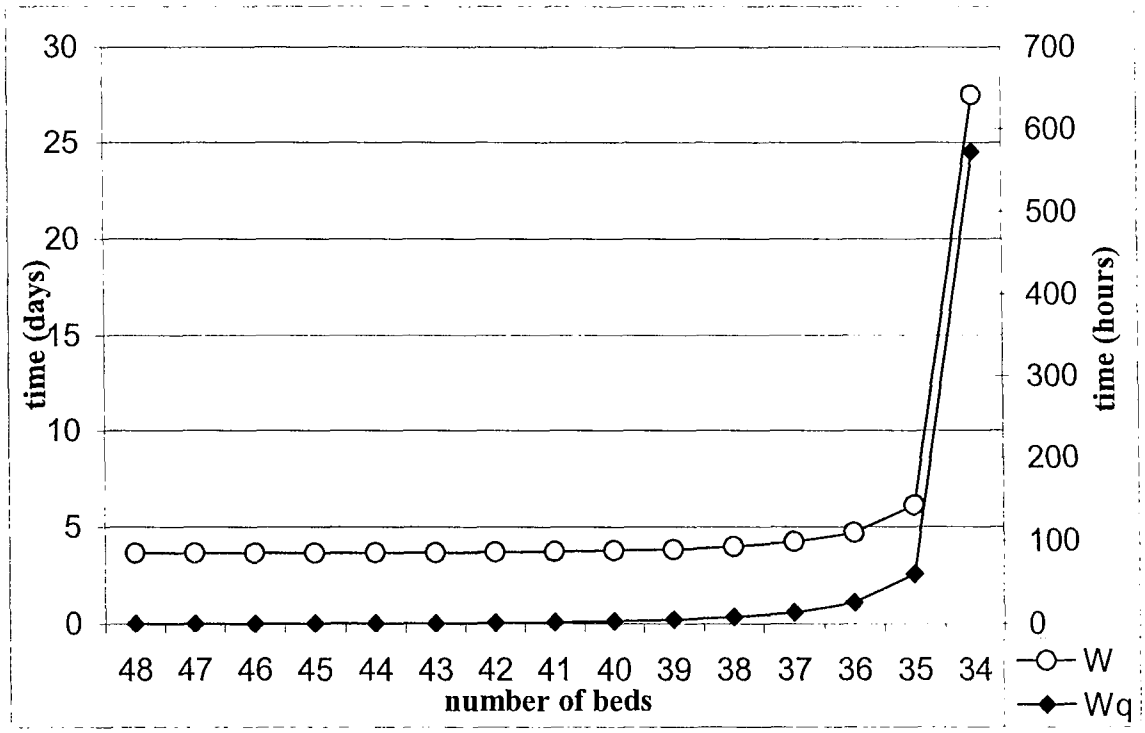


Figure 11 Average time in system and in queue when number of beds decreases.



Conclusion

Even though healthcare in the United States is very costly and hospitals have a tremendous problem with waiting patient time, they are still reluctant to use queueing theory to resolve their waiting problem by efficient bed resource allocation using queuing theory. This part of the thesis uses queueing theory in medical units of Sentara Leigh Hospital, in Norfolk, Virginia. The queuing model provided the insights to see the trade-off between utilization (system perspective) and patient waiting time (customer perspective). Created different scenarios will enable hospital managers to see the effect of arrival rate, service rate, and number of beds to estimate the main performance

measures of assessing the benefits of providing extra beds to minimize patient waiting time when demand increases.

Future research

1. The queuing model results were based on only 33 days. More data is needed to validate the queuing model.
2. Estimate arrival and service time by considering seasonality. If there is seasonality related to hospital's admissions, then hospital managers from the queuing model will see how to reflect on this seasonality since the number of beds and utilization will change.
3. Simulation. Develop a simulation model will be beneficial to further validate our model and to conduct experiments that cannot be handled by queuing models.

CHAPTER V

V. ESTIMATING PROCEDURE TIME

The second part of this thesis examines a method of estimating procedure times in surgical operating rooms. This part consists of four sections: Section One provides an introduction; Section Two reviews the literature; Section Three provides a methodology; and, finally, Section Four provides the results, conclusion, and recommendations for future research.

INTRODUCTION

Surgical operating rooms are one of the most costly functional areas in hospitals (Epstein, 1995, Redelmeier and Fuchs, 1993, American Hospital Association, 1994). One of the reasons for such high costs is inefficient scheduling. In their work, Litvak et al. (2000) provided evidence that scheduling in practice is often inefficient. A number of different approaches have been applied to the problem of scheduling surgical cases, such as mathematical programming, computer-based simulation, and algorithm evaluation (Dexter et al, 1999); optimization techniques (Blake et al., 2002, Blake and Donald, 2002); rule-based heuristic approaches (Dexter and Macario, 2002); and statistical decision theory (Dexter and Traub, 2000).

But before hospital managers spend time on improving scheduling in their hospitals, they need to look at the root cause of the problem, which requires finding a better way to estimate surgical procedure times. If surgical procedure times are inaccurate, then consequently scheduling will be inefficient. Hospital senior managers and administrators can reduce costs incurred in inefficient scheduling by better estimating surgical cases (procedures). Inaccurate estimation of procedure time increases costs when overestimation results in unused operating rooms, and underestimation results in overtime and cancellation of procedures. Indeed, such inaccuracy leads not only to increasing costs but also to the dissatisfaction of patients, surgeons, and operating room staff. While in theory it is simple enough to claim that more accurate methods of estimating procedure times are necessary, in practice efficient surgical scheduling can be complicated by a variety of problems. It can be complicated by several factors including the variability factor inherent in the duration of surgical procedures, or by a particular doctor and his/her experience, as well as by particular procedures and patient types.

Determining an appropriate statistical model for surgical procedure times is very important for several reasons. First, it is important in order to identify the particular surgeons and procedures that contribute the most to variability. An appropriate model will identify atypically slow or fast-working surgeons, or certain outliers that can be scheduled separately. Second, it is important in order to estimate surgical procedure times based in a statistical distribution. The most appropriate statistical model and its accurate parameters are crucial to scheduling. Relying on standard distribution models in cases where another distribution would be more appropriate can lead to false statistical results and therefore, to inefficient scheduling.

LITERATURE REVIEW

Hospital managers and administrators have been interested in efficient modeling of surgical procedures for at least 40 years. Rossiter and Reynolds (1963) showed that waiting times appear as lognormal distribution. Zhou and Dexter (1998) estimated the prediction bounds for the duration of the next surgical cases based on the assumption that lognormal distribution indeed fit the data. Strum et al. (2000) showed that the appropriate statistical distribution for surgical procedure times is lognormal.

Zhou and Dexter (1998) calculated the "upper prediction bound," which specified the duration of the next procedure time will be less or equal to a certain probability bound. They estimated the prediction bounds using two methods: First, by assuming procedure times are distribution free, and second, by assuming procedure times follow a lognormal distribution. Procedure times have two characteristics of a lognormal distribution: First, they have positive values, and second, only small number of procedure times may take longer than the average. The results from the study showed that prediction bounds could be accurately estimated assuming that procedure times are lognormal.

Recently, Strum et al. (2000) indicated that surgical procedure times appear more as lognormal distribution rather than normal. The authors tested whether the distribution of surgical procedures times fit more closely lognormal or normal distribution by using different statistical methods. They analyzed total procedure time as the time from entry into the operating room until emergence from anesthesia, and surgical procedure time as

the time from the initial incision to closure of the surgical wound. They used case frequencies of five or more procedures, enough to fit the probability distribution. The data was subdivided into subgroups according to Current Procedural Terminology (CPT) in combination with anesthesia type, as opposed to CPT alone. The overall performance of the lognormal and normal distribution models was compared and the results showed that both surgical procedure times and total procedure times fit well with the lognormal model distribution.

Later Strum et al. (2003) tested dual procedure surgeries performed in the same surgical session to determine if the lognormal distribution was superior to the normal distribution. The results showed that the lognormal models fit better than normal model. The authors suggested that it might be practical to estimate the next dual procedure by considering the duration of the longest procedure and the type of anesthesia.

Most of the time hospital managers and senior administrators simply assume that surgical times are normally distributed, estimating surgical time by using mean and plus/minus standard deviations (Strum et al. 2000). This is precisely what we found in the case of Sentara Leigh Hospital's method of estimating procedure time. Case duration is estimated by taking the mean of a surgeon's historical data for the last 10 procedures and assuming that procedures follow a normal distribution. High and low values are dropped. Set up and preparation times are estimated according to the type of procedure. Then, the average of 8 procedures, plus setup and preparation types, is used to estimate the time needed to finish the procedures.

Judging from the literature review, it can be concluded that presuming procedure time normally distributed can lead to false or erroneous results. Therefore, instead of

assuming that surgical procedure times follow the normal distribution and taking only the average of procedure times, a more appropriate statistical distribution is needed. In this part of this thesis, a more effective methodology to estimate surgical procedure times is proposed.

PROPOSED METHODOLOGY

3166 surgical procedures from Sentara hospital performed in the main operating room (MOR) during a 6-month period starting from March 2004 and ending August 2004 were studied. The total time was defined as the time of entry into the operating room until the time of departing the operating room. The total procedure time was defined as the time when the procedure started until the time of completion. The data of procedure times was subdivided into homogeneous groups, according to particular surgeons performing a certain case. The reason for this is to identify surgeons that contribute the most to the difference between the actual and estimated procedure times, and what distribution they fit.

The cases performed by each surgeon less than 8 times during those 6 months were omitted since the sample size will be too small to represent the population. To show that there was a significant difference between the estimation of procedure times and actual procedure times, a *t* test was performed. After the *t* test, only 642 cases were left, which are equal to 20.3 % of the total number of cases. Using Pareto's 80/20 Rule, it is not unusual that 20 percent of the reasons cause 80 percent of the problems. Managers know that 20 percent of the cases consume 80 percent of time and resources. From our results, it can be concluded that 20 percent of surgical procedures causes 80 percent of

over- or underestimation of surgical cases. Thus, in our analysis, we studied the 20 percent of procedures.

So as to inspect what distribution type the surgical procedures follow, the Input Analyzer in Arena simulation software was employed. As a result of this application, it appears that a total of 642 procedures were consistent with a lognormal distribution. The p values of lognormal distribution fitting are large and are presented in Appendix B.

The lognormal distribution was chosen based on what has been cited in the literature review and its applicability to the problem. Lognormal distribution takes values from zero to infinity and is skewed to the right. Since the procedure times were consistent with lognormal distribution, instead of taking the mean of a normal distribution, the mode of a lognormal distribution was taken as an estimate for the procedure time. The mode was chosen as an estimate because it provides the measurements that occur most frequently in the data set. Since the mean is the average of a data set and can be influenced by extreme values, in this case we decided the mode is a better representative of central tendency than the mean.

RESULTS

The results and histograms presented in this section are only for two surgeons (Surgeon A and Surgeon B). The rest of the histograms pertaining to the other surgeons can be found in Appendix B.

Surgeon A performed 20 procedures types CYCKUB over the sixth month period. Surgeon B performed 66 procedures types GEGASTB over the same sixth month period.

Figure 12 Histogram for surgeon A, procedure type: CYCKUB.

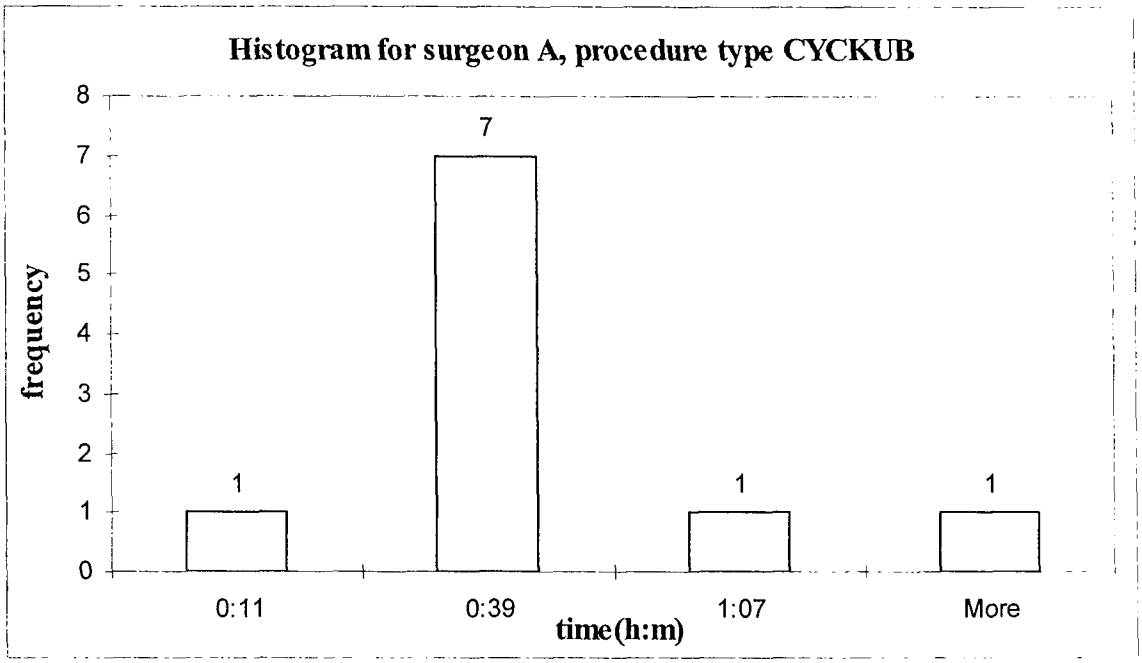
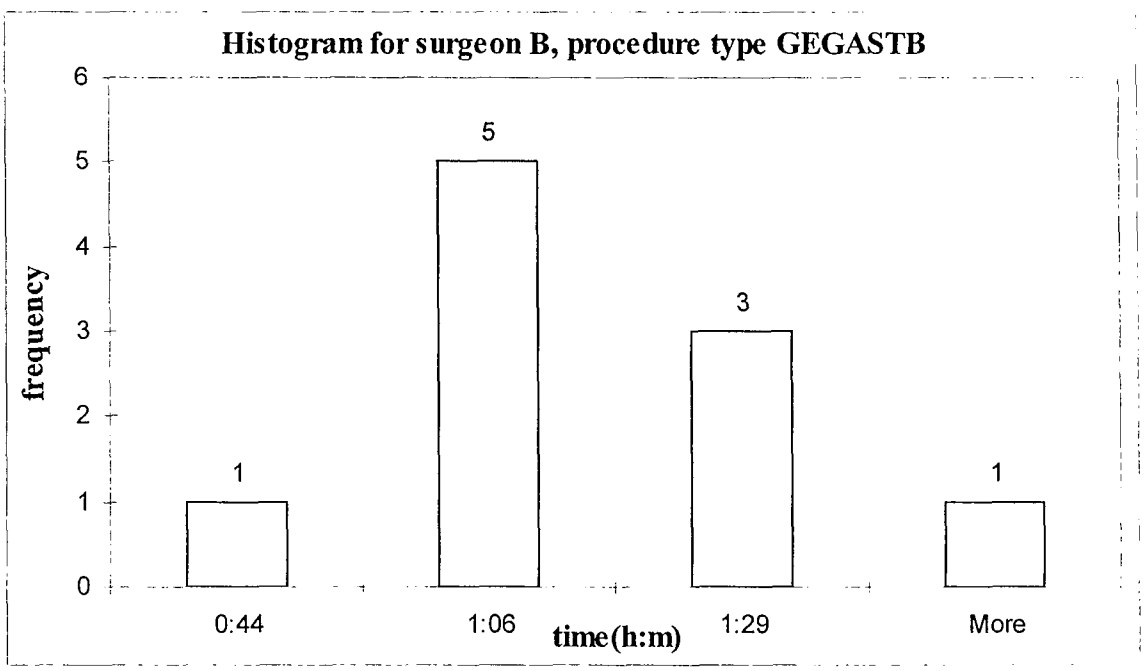


Figure 13 Histogram for surgeon B, procedure type: GEGASTB.



From the histograms shown in Fig.1 and 2, the most frequent surgical time was used as an estimate for the next procedure times. In the case of Surgeon A, the estimated time for subsequent procedures was forty minutes. The scheduled time of the current method was varied from thirty minutes to one hour and thirty minutes over sixth month period. In the case of Surgeon B, the estimated time for the subsequent procedures was one hour and five minutes. The scheduled time of the current method was one hour.

In the following table, the "Current" column shows that the difference between the actual performed time and current scheduled time is 4:29 hours for Surgeon A and 11:06 for Surgeon B. The "Proposed" column shows that the difference between the actual performed time and proposed time (highest peak) is 3:07 for Surgeon A, and 10:00 hours for Surgeon B. In the "Difference" column, the percentage numbers show that the proposed method is better than the current method by 30.48 % in the case of Surgeon A and by 9.91 % in the case of the Surgeon B.

Table 8 Summary of the results for surgeon A and B.

# of procedures	Surgeon's name	Procedure type	p value	Current	Proposed	Difference
20	Surgeon A	CYCKUB	> 0.15	4:29	3:07	30.48%
66	Surgeon B	GEGASTB	0.05	11:06	10:00	9.91%

A summary of the results for all studied surgeons is in Appendix B.

CONCLUSION

Of all cases performed during the sixth month period, 20 percent contributed to the significant difference between the actual and scheduled time. Therefore, the current method of estimating procedure times based on the assumption that procedures occur according to a normal distribution is not effective. In this section of the thesis, a more effective methodology to estimate surgical procedure times has been proposed. Since an appropriate distribution is very important in estimating procedure times, the data was fit to determine which statistical distribution the procedures follow. It appeared that procedure times were consistent with a lognormal distribution. After creating the histograms for each surgeon performing a certain type of procedure, the highest peak value of lognormal distribution (mode) was chosen for estimating the next procedure times. The results of the 20 percent procedures from the proposed method in comparison with the current method clearly indicate that better results can be achieved with the former. The difference between the actual and proposed procedure time was less than that between the actual time and current scheduled time. The proposed methodology is efficient not only in estimating the next procedure time every time but also in estimating future procedure times over a six-month horizon at least. Extending the scheduling horizon will save time and money. Moreover, it may improve satisfaction rates among patients and hospital staff by establishing a framework capable of providing more realistic expectations for procedural and wait times.

FUTURE RESEARCH

We propose two avenues of future research based on the findings of this study. First, even though the *t* test showed a significant difference between the actual and scheduled time in 20 % of all cases, it will be necessary to examine the cases which did not show a significant difference to see which distribution they follow. The next procedure time needs to be estimated on the results of an appropriate distribution rather than on the erroneous assumption that the distribution procedures will always follow a normal distribution.

Second, the use of regression analysis can be employed to arrive at better estimates of procedure time. In order to accomplish that, one needs to identify the relevant factors that may contribute to delays in procedure time. For example, doctor's experience, patient age, equipment specification among others. To do this separate study, the necessary historical data must be launched.

REFERENCES

American Hospital Association (1994). Ambulatory Surgery Trendlines. Chicago: Society for Ambulatory Care Professionals.

Bailey, N.T.J. (1952). A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society, Series B, 14,* 185-199.

Blake, J., & Donald, J. (2002). Mount Sinai hospital uses integer programming to allocate operating room time. *Interfaces 32,* 66-73.

Blake, J., Dexter, F., & Donald, J. (2000). Operating room manager's use of integer programming for assigning block time to surgical groups: A case study. *Anesthesia and Analgesia 94,* 143-148.

Bunday, D. (1986). *Basic Queueing Theory.* London: Edward Arnold Ltd.

Cohen, M.A., Hershey, J.C., & Weiss, E.N. (1980). Analysis of Capacity Decisions for Progressive Patient Care Hospital Facilities. *Health Services Research 15,* 145-160.

Dexter, F. & Traub, R. (2000). Sequencing cases in the operating room: Predicting whether one surgical case will last longer than another. *Anesthesia and Analgesia 90,* 975-979.

Dexter, F., & Macario A. (2002). Changing allocations of operating room time from a system based on historical utilization to one where the aim is to schedule as many surgical cases as possible. *Anesthesia and Analgesia 94,* 1272-1279.

Dexter, F., Macario, A., & Traub R. (1999). Which algorithm for scheduling add-ion elective cases maximizes operating room utilization? Use of bin packing algorithms and fuzzy constraints in operating room management. *Anesthesiology 91,* 1491-1450.

El-Darzi, E., Vasilakis, C., Chaussalet, T., & Millard, P. (1998). A simulation modeling approach to evaluating length of stay, occupancy, emptiness, and bed blocking in a hospital geriatric department. *Health Care Management Science 1,* 143-149.

Epstein, A. (1995). US teaching hospital in the evolving health care systems. *JAMA 273,* 1203-1207.

Esogbue, A., & Singh, A. (1976). A Stochastic Model for an Optimal Priority Bed Distribution Problem in a Hospital Ward. *Operations Research 24,* 884-895.

Gorunescu, F., McClean, S., & Millard P. (2002). Using a Queuing Model to Help Plan Bed Allocation in a Department of Geriatric Medicine. *Health Care Management Science 5,* 307-312.

Green, L.V., & Nguyen V. (2001). Strategies for Cutting hospital Beds: The Impact on Patient Service-Statistical data included. *Health Services Research 36,* 421-442.

Green, L.V. (2002). How Many Hospital Beds? *Inquiry 39,* 400-412.

Gross, D., & Harris, C.M. (1985). *Fundamentals of Queuing Theory.* New-York: John Wiley & Sons.

Hall, R. (1991). *Queuing methods: for services and manufacturing.* New Jersey: Prentice Hall.

Kim, S., Horowitz, I., Young, K., & Buckley, T. (1999). Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research 115,* 36-46.

Litvak, E. & Long, M.C. (2000). Cost and quality under managed care: Irreconcilable differences? *The American Journal of Managed Care 6 (3),* 305-312.

McClain, J.O. (1976). Bed planning Using Queueing Theory Models of Hospital Occupancy: A Sensitivity Analysis. *Inquiry 13,* 167.

McClave, J., Benson, G., & Sincich, T. (1998). *Statistics for business and economics.* New Jersey: Prentice Hall.

McClure, W. (1976). *Reducing Excess Hospital capacity.* Excelsior, Minn.: Bureau of health planning.

McManus, M.L., Long, M.C., Cooper, A., Mandel, J., Berwick, D.M., Pagano, M., & Litvak, E. (2003). Queuing Theory Accurately Models the Need for Critical Care Resources. *Anesthesiology 98,* 1491-1496.

Redelmeier, D., & Fuchs, V. (1993). Hospital expenditures in the United States and Canada. *New England Journal of Medicine 328,* 772-778.

Rossiter, C., & Reynolds, J. (1963). Automatic monitoring of the time waited in out-patient departments. *Medical Care 1,* 218-225.

Saaty, T. (1961). *Elements of Queuing Theory.* York, PA: Maple Press Company.

Shonick, W. (1970). A Stochastic Model for Occupancy-Related Random Variation in General-Acute Hospitals. J. *Am. Stat. Assoc. 65,* 1974.

Shonick, W., & Jackson, J.R. (1973). An Improved Stochastic Model for Occupancy Related Random variables in General-Acute Hospitals. *Operations Research 21,* 952.

Siddharthan K., Walter, J.J., & Johnson, J. (1996). A priority queuing model to reduce waiting times in emergency care. *International Journal of Health Care Quality Assurance 9/5,* 10-15.

Spangler, W.E., Strum, D., Vargas, L., & May, J., (2004). Estimating Procedure times for Surgeries by determining Location parameters for the Lognormal Model. *Healthcare Management Science 7,* 97-104.

Spatz, C., & Johnston, J., (1981). *Basic Statistics: Tales of distributions.* Monterey, CA: Brooks/Cole Publishing Company.

Strum, D., May, J., Sampson, A., Vargas, L., & Spangler, W. (2003). Estimating Times of Surgeries with two component procedures: Comparison of Lognormal and Normal models. *Anesthesiology 98,* 232-240.

Strum, D., May, J., & Vargas, L. (2000). Modeling the uncertainty of surgical procedure times. *Anesthesiology 92,* 1160-1167.

Tucker, J.B., Barone, J.E., Cecere, J., Blabey, R.G., & Rha, C.K. (1999). Using Queuing theory to Determine Operating room Staffing Needs. *J Trauma 46,* 71-79.

Weiss, E., & McClain, J.O. (1987). Administrative Days in Acute Care Facilities: a Queuing-Analytic Approach. *Operations Research 35,* 35-44.

Welch, J.D. (1964). Appointment systems in hospital outpatient departments. *Operational Research Quarterly 15,* 224-237.

Worthington, D.J., (1987). Queuing Models for Hospital Waiting lists. *Journal of Operational Research Society 38 (5),* 413-422.

Zhou, J., & Dexter, F. (1998). Method to Assist in the Scheduling of Add-on Surgical cases- Upper Prediction bounds for Surgical Case Durations Based on the Lognormal Distribution. *Anesthesiology 89(5),* 1228-1232.

APPENDIXES

APPENDIX A

Performance measures when arrival rate 10 patients/24h.

Average server utilization	88	%
Average number of customers in the queue (L_q)	2.83	patients
Average number of customers in the system (L)	39.22	patients
Average waiting time in the queue (W_q)	6:46	hour/patient
Average time in the system (W)	94:07	hour/patient

Performance measures when arrival rate 9 patients/24h.

Average server utilization	79.9	%
Average number of customers in the queue (L_q)	0.46	patients
Average number of customers in the system (L)	33.22	patients
Average waiting time in the queue (W_q)	1:13	hour/patient
Average time in the system (W)	88:35	hour/patient

Performance measures when arrival rate 8 patients/24h.

Average server utilization	71	%
Average number of customers in the queue (L_q)	0.06	patients
Average number of customers in the system (L)	29.18	patients
Average waiting time in the queue (W_q)	0:11	hour/patient
Average time in the system (W)	87:32	hour/patient

Performance measures when service rate 3 days/patient.

Average server utilization	68.05	%
Average number of customers in the queue (L_q)	0.03	patients
Average number of customers in the system (L)	27.93	patients
Average waiting time in the queue (W_q)	0:04	hour/patient
Average time in the system (W)	72:04	hour/patient

Performance measures when service rate 3.5 days/patient.

Average server utilization	79.39	%
Average number of customers in the queue (L_q)	0.41	patients
Average number of customers in the system (L)	32.96	patients
Average waiting time in the queue (W_q)	1:03	hour/patient
Average time in the system (W)	85:03	hour/patient

Performance measures when service rate 4 days/patient.

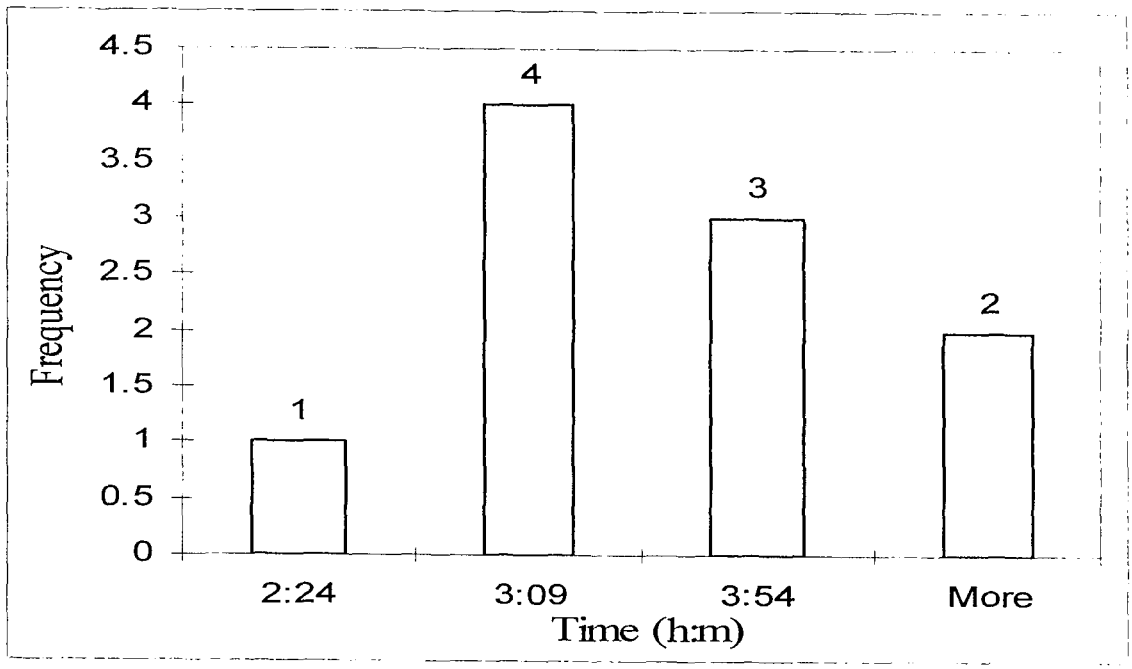
Average server utilization	90.73	%
Average number of customers in the queue (L_q)	4.29	patients
Average number of customers in the system (L)	41.29	patients
Average waiting time in the queue (W_q)	11:04	hour/patient
Average time in the system (W)	107:04	hour/patient

APPENDIX B

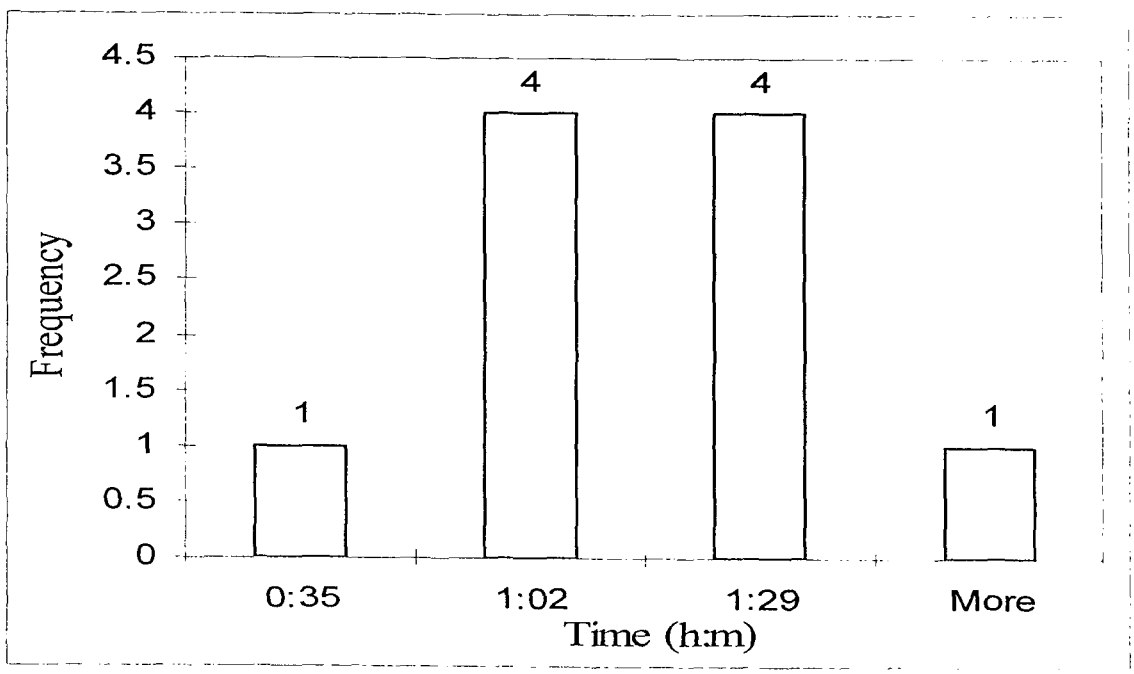
P values from lognormal distribution fitting.

Number of procedures	Surgeon's name	Procedure type	P value
8	Surgeon L	ORLUMLAM	> 0.15
8	Surgeon N	ORARTKN	> 0.15
12	Surgeon O	GELAPGAS	> 0.15
8	Surgeon P	GEGASTB	> 0.15
16	Surgeon A	ORLUMLAF	> 0.15
18	Surgeon B	GYSACFX	> 0.15
10	Surgeon D	PLBRER1	> 0.15
102	Surgeon C	ORTOTKRP	< 0.01
34	Surgeon C	ORARTKN	< 0.01
9	Surgeon C	ORTOTKRV	> 0.15
9	Surgeon R	DEMANDOS	> 0.15
10	Surgeon S	PLBRERD	> 0.15
13	Surgeon T	NEANTCVF	> 0.15
10	Surgeon Y	CYCKUB	> 0.15
37	Surgeon Z	ORTOTHRP	0.01
8	Surgeon S	ORLUMDEF	0.11
10	Surgeon W	ORANKFO	> 0.15
12	Surgeon V	GYTAH	> 0.15
17	Surgeon E	GYEXPLAP	> 0.15
20	Surgeon F	CYCKUB	> 0.15
104	Surgeon G	GEGASTB	< 0.01
43	Surgeon K	ORTOTHRP	> 0.15
23	Surgeon K	ORROTCU	0.03
27	Surgeon K	ORARTKN	0.02
8	Surgeon K	ORTOTSRP	> 0.15
66	Surgeon M	GEGASTB	0.05

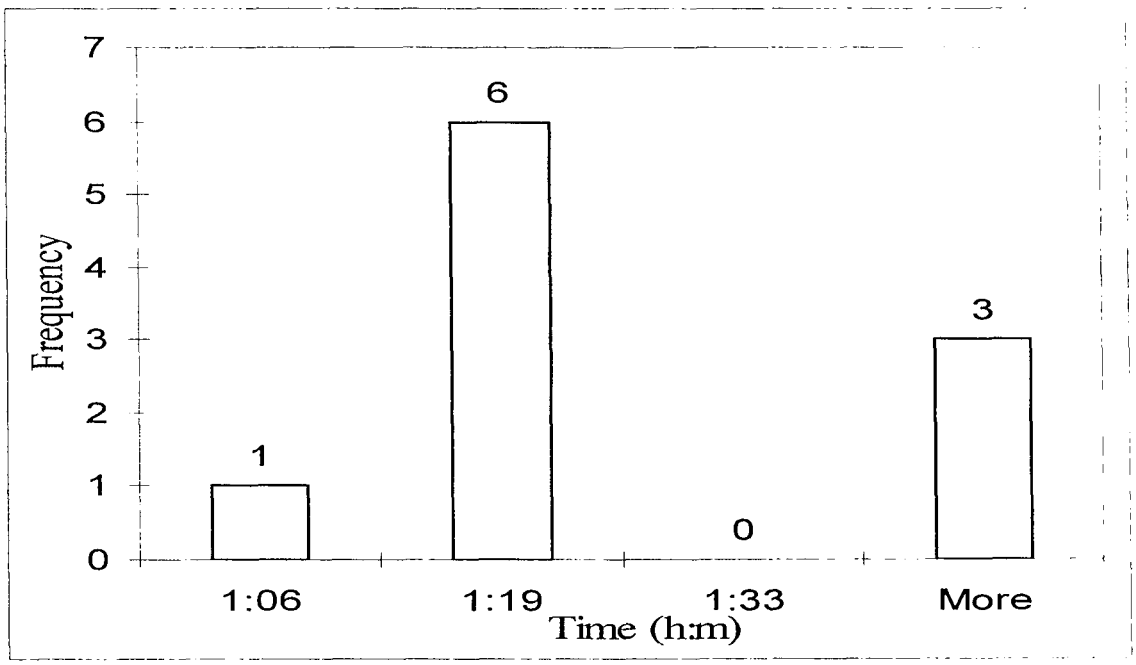
Histogram for Surgeon A, procedure type: ORUMLAF.



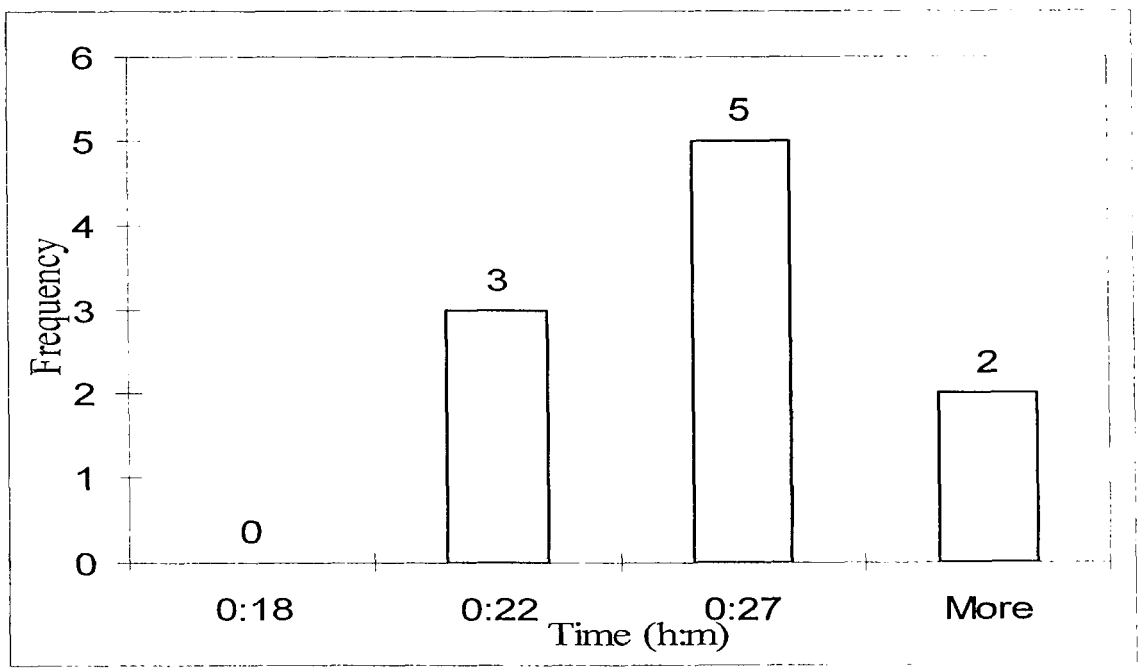
Histogram for Surgeon B, procedure type: GYSACFX.



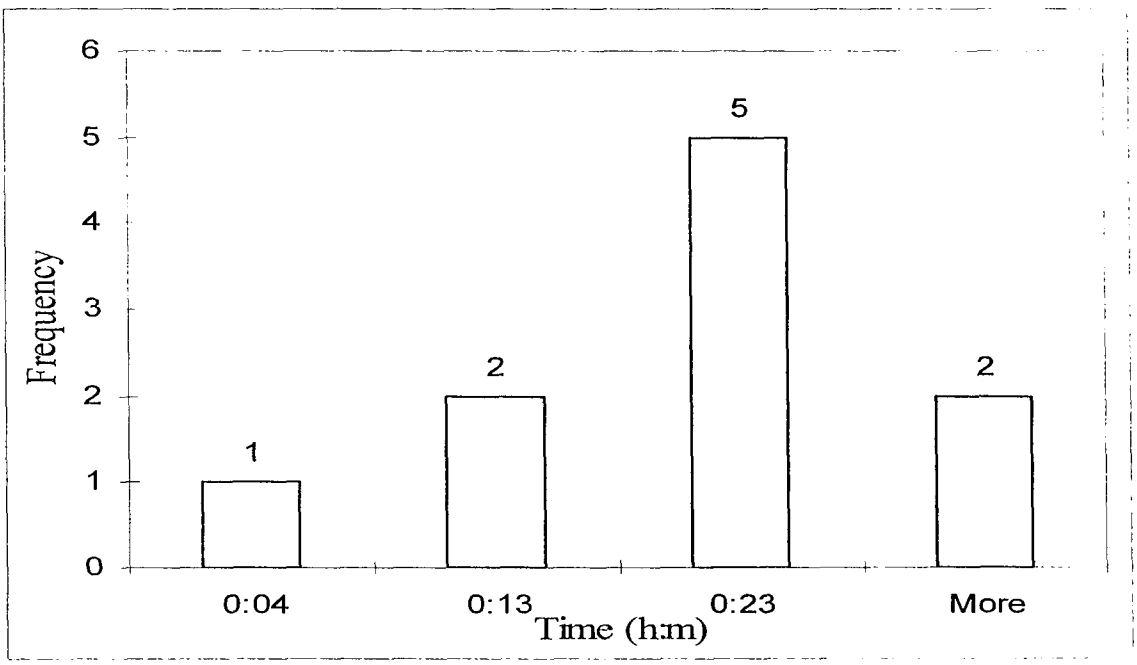
Histogram for Surgeon C, procedure type: ORTOTKRP.



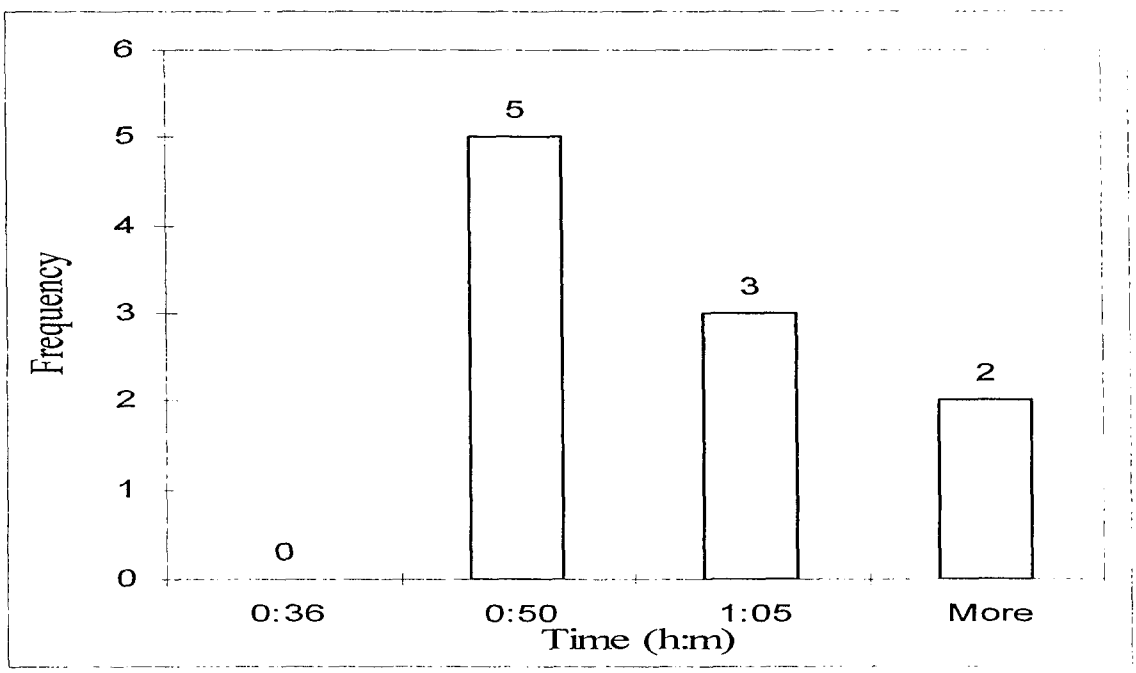
Histogram for Surgeon C, procedure type: ORARTKN.



Histogram for Surgeon K, procedure type: ORARTKN.



Histogram for Surgeon K, procedure type: GEGASTB.



Summary of the results for all surgeons.

Number of procedures	Surgeon's name	Procedure type	Current	Proposed	Difference between current and proposed method
16	Surgeon A	ORLUMLAF	7:35	7:37	0.44%
18	Surgeon B	GYSACFX	3:37	2:07	41.50%
102	Surgeon C	ORTOTKRP	3:20	0:31	10.00%
34	Surgeon C	ORARTKN	2:58	2:19	21.91%
37	Surgeon D	ORTOTHRP	8:47	5:53	33.02%
17	Surgeon E	GYEXPLAP	4:45	4:01	15.44%
20	Surgeon A	CYCKUB	4:29	3:07	30.48%
104	Surgeon G	GEGASTB	18:31	13:02	29.61%
43	Surgeon K	ORTOTHRP	11:02	10:31	4.68%
23	Surgeon K	ORROTCU	4:00	2:29	37.92%
27	Surgeon K	ORARTKN	3:29	2:49	19.14%
66	Surgeon B	GEGASTB	11:06	10:00	9.91%