UTILITY ANALYSIS OF AN INTENSIVE CARE UNIT MEDICAL SERVICE MODEL USING QUEUING THEORY WITH IMPROVED TAGUCHI LOSS FUNCTION

BY

KIPRONO DAVID KIPKORIR

(MT. KENYA UNIVERSITY, BED SCIE)

A Thesis Submitted to the School of Post-Graduate Studies in Partial Fulfillment of the Requirements of the Degree of Master of Science in Applied Mathematics of the School of Pure and Applied Sciences

Kisii University

October, 2017

DECLARATION

Declaration by the Candidate

I the undersigned declare that this is my original work and has not been presented for academic purposes in this University or any other University whatsoever.

Sign……………………………………………. Date…………………………………

Name: David Kiprono

MPS12/70036/14

Recommendation by the Supervisors

This thesis report has been submitted for examination with our approval as University supervisors.

Sign………………………………………… Date…………………………………

Dr. Titus Rotich, PhD

Department of Centre for Teacher Education, (Science Education)

Moi University.

Sign………………………………………… Date………………………………...

Dr. Joash M. Kerongo, PhD

Department of Mathematics and Actuarial Sciences

P.O Box 408- 40200, Kisii.

Kisii University

PLAGIARISM DECLARATION

Definition of plagiarism

Is academic dishonesty which involves; taking and using the thoughts, writings, and inventions of another person as one's own.

DECLARATION BY STUDENT

i. I declare I have read and understood Kisii University Postgraduate Examination Rules and Regulations, and other documents concerning academic dishonesty.
ii. I do understand that ignorance of these rules and regulations is not an excuse for a violation of the said rules.
iii. If I have any questions or doubts, I realize that it is my responsibility to keep seeking an answer until I understand.
iv. I understand I must do my own work.
v. I also understand that if I commit any act of academic dishonesty like plagiarism, my thesis/project can be assigned a fail grade ("F")
vi. I further understand I may be suspended or expelled from the University for Academic Dishonesty.

Name__________________________ Signature_____________________

Reg. No____________________________ Date_________________________

DECLARATION BY SUPERVISOR (S)

i. I/we declare that this thesis/project has been submitted to plagiarism detection service.
ii. The thesis/project contains less than 20% of plagiarized work.
iii. I/we hereby give consent for marking.

1. Name__________________________ Signature_____________________

   Affiliation ________________________ Date_________________________

2. Name__________________________ Signature_____________________

   Affiliation ___________________________ Date_________________________

DECLARATION OF NUMBER OF WORDS

DECLARATION OF NUMBER OF WORDS FOR MASTERS THESES

This form should be signed by the candidate and the candidate's supervisor (s) and returned to Director of Postgraduate Studies at the same time as you copies of your thesis/project.

Please note at Kisii University Masters and PhD thesis shall comprise a piece of scholarly writing of not more than 20,000 words for the Masters degree and 50 000 words for the PhD degree. In both cases this length includes references, but excludes the bibliography and any appendices.
Where a candidate wishes to exceed or reduce the word limit for a thesis specified in the regulations, the candidate must enquire with the Director of Postgraduate about the procedures to be followed. Any such enquiries must be made at least 2 months before the submission of the thesis.
Please note in cases where students exceed/reduce the prescribed word limit set out, Director of Postgraduate may refer the thesis for resubmission requiring it to be shortened or lengthened.

Name of Candidate: …………………………………… ADM NO…………………

Faculty…………………………………… Department…………………………….......

Thesis Title: …………………………………………………………………….............

………………………………………………………………………………………

………………………………………………………………………………………

I confirm that the word length of:

1) the thesis, including footnotes, is ……… 2), the bibliography is ……… and, if applicable, 3) the appendices are ………

I also declare the electronic version is identical to the final, hard bound copy of the thesis and corresponds with those on which the examiners based their recommendation for the award of the degree.

Signed: ……………………………………………………. Date: ……………………
David Kiprono

I confirm that the thesis submitted by the above-named candidate complies with the relevant word length specified in the School of Postgraduate and Commission of University Education regulations for the Masters and PhD Degrees.

Signed: .... Email………….……………………Tel……………. Date: ……….
Dr. Titus Rotich, PhD

Signed: .... Email………….……………………Tel……………. Date: ……….
Dr. Joash M. Kerongo, PhD

COPYRIGHT

No part of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or means such as electronic, mechanical, photocopying, recording without prior written permission from the author and/or Kisii University or her behalf.

© 2017, Kiprono David

DEDICATION

I dedicate this work to my teachers who showed me the way, my family members for encouraging me walk the way and my colleagues who gave me the morale to persist until the end. I thank them for all their support that they gave me to do the work. May almighty God bless them abundantly.

ACKNOWLEDGEMENT

I thank the Almighty God who makes everything possible. I am grateful to everyone who contributed in this research especially close friends who encouraged me all the way and gave me strength, knowledge and sound mind to carry out the task. Also my gratitude goes to my dear wife, and my children for their prayers and financial support they gave unto me also to my supervisors Dr. Kerongo Joash and Dr. Rotich Titus for their continuous guidance. Lastly I thank the management of Moi Teaching and Referral Hospital for the assistance they accorded to me to get the data required for this work.

ABSTRACT

A common situation that occurs in everyday life is that of queuing or waiting in the line for services. Long queues have become a major source of concern in all service facilities and the most affected are the Intensive Care Units in medical facilities. This study is therefore a utility analysis of queuing problem at Moi Teaching and Referral Hospital (MTRH) Intensive Care Unit (ICU) in Kenya. The objectives were to determine the average time of a patient in the system, optimum number of beds required and establish the stability of the system using time and costs of the system. Admission data of ICU for six months was obtained from MTRH. Due to the nature of the problem, a Multi-server queuing Model (M/M/s) was used together with Improved Taguchi Loss Function to analyze the problem and an excel calculator was used to simulate the model results in five scenarios. It was found that the optimum number of beds required in the ICU was 13, which reduces the patient waiting time by 86.06% while server utilization remains good at 77%. Lastly, the stability of the system was found out to be achieved when the bed allocation is between 12 and 14 by using the total expected costs together with improved Taguchi Loss Function. Therefore, from the findings of this work, it is recommended that MTRH management, policy makers at county and national level and other health facilities with similar queuing problem improve the overall patient care by installing the optimum number of beds in order to meet the patient needs. The significance of the study is to provide sufficient information to the health service providers, county governments and national governments improve service delivery to reduce customer mortality rate.

TABLE OF CONTENTS

CHAPTER FIVE

LIST OF ABBREVIATIONS AND ACRONYMS

ED	=	Emergency Department
EMS	=	Emergency Medical Service
ESC	=	Expected Service Cost
ETC	=	Expected Total Cost
EWC	=	Expected Waiting Cost
FIFO	=	First In First Out.
ICU	=	Intensive Care Unit
LSL	=	Lower Specification Limit
MTRH	=	Moi Teaching and Referral Hospital
OT	=	Operation Theatre
TLF	=	Taguchi Loss Function
USL	=	Upper Specification Limit

LIST OF SYMBOLS

P_0 = probability of 0 customers in system

P_n = probability of exactly n customers in system

W = average time in the system

W_q = average waiting time

λ = average number of arrivals per unit time period

μ = average number of people served per unit time.

ρ = utilization factor $\rho = {}^{\lambda}/_{\mu}$

T = target value or quality of service

R = Cost of rejection at the specification limit.

L = average number in the system

Lq = average number in the queue

C_s = Service Cost

C_w = Waiting Cost

C_T = Total Cost

s = the number of servers,

M = Inter arrival time and inter service time

L_q = Expected number of customers in the queue,

L_s = Expected number of customers in the system,

λdt =probability that an arrival enters the system between t and $t + dt$ time interval within time interval dt.

$1 - \lambda dt$ =probability that no arrival enters the system within interval dt plus higher order terms

$\mu dt =$ Probability of one service completion between t and $t + dt$ time interval within time interval dt.

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Waiting to receive service in a queue happen everywhere, it affects people in polling stations as they queue to vote, traffic on the road, patients in hospital, customers in shops, buying fuel from a petrol station, queuing on the bank Automatic Teller Machine (ATM) to withdraw cash, or making withdrawals or deposits in a bank that still require customers to queue physically. Though, currently we must appreciate the automated customer queuing in most banks in Kenya where customers get their number in relation to the type of service required and simply wait for their turn to be announced. This changes the bank scenario to a Multi-Phase, Multi Server queuing system. These queues are people lining up to be served or are machines that are waiting to be repaired, Lorries lining to be loaded or unloaded, or aeroplanes waiting to take land or take off in a airport. (Resing, Adan and Jacques, 2015) Wrote that queuing models have many important application areas and some that he stated include; medical services, communication services, production lines, transportation networks, stocking and information processing services. Queuing models are mainly useful in the planning of these services in terms of layout, capacity and control.

Foster., Michael & Ziya, (2010) observed that Queuing models provide solutions to problems of people waiting to receive service. That is why they are also particularly relevant in health care. Generally, they showed the applicability of modelling in health care service delivery.

The first application of queuing theory, which is in fact the one that stimulated the development of the whole research area, was the design and analysis of telephone network. In 20th century, in the early years, operators received telephone calls first before they connected to the person that the call was intended for, (Baun and Breuer 2005). He further states that, because of the diminishing demand of telephone calls due to the huge traffic, Erlang was tasked to provide a solution to the problem. He carried out experiments and later came up with a report on how to address delays problem in telephone calls automatic dialling. Success of his work encouraged the use of queuing in many other queuing problems.

A queuing process has arrivals, service points, and customers waiting in line to be attended to by the service provider. The cost of ensuring quality, in the provision of products and services, is something that is difficult to measure. For example, the length of time a customer will wait in line before being served is a key measure of the quality perceived, and therefore a contributing factor of customer satisfaction, (Kembe, Onah and Iorkegh, 2012).

According to Aronsky (2008), Emergency Department overcrowding is an international crisis that affects the quality of health care service. This is indeed true in that most hospitals offering Emergency services are overcrowded and many patients will not receive the service in time or may even loose life waiting.

A study conducted by Paul and Li(2008) described that a hospital's efficiency in service provision depends on the number of available staff, availability of intact medical equipment's including operating rooms, laboratory, supplies of water, power, medical gases like oxygen and the state of the building. This scenario is evident in many hospitals which do not have enough facilities to handle all the incoming traffic.

Obamiro (2010) studied the waiting line for expectant women in Ante natal care unit. The results of the study evaluated the effectiveness of a queuing model in identifying the shortcomings in the facilities that served the expectant mothers. The greatest challenge was the waiting time of each mother to receive service depending on her time of arrival.

Schoenmeyr, Dunn and Gamarnik, (2009) analysed some of the healthcare organizations functioning with very small differences, so decisions on compelling the scarce resources must be done well so that the investment will lead to the desired result. Queuing approach to waiting time problems is useful because it enables the research of future scenarios for which historical data are not applicable. Waiting times calculations assist in establishing the rate of service on hospital waiting lists and are a more reliable measure of hospital performance than the size of the waiting list. In some cases the patient may be removed from a waiting list and the reasons may include that they no longer require the procedure, are instead admitted as an emergency patient, receive their treatment at a different hospital.

1.2 Statement of the Problem

Successful service providing entities strive to provide the best services to their customers while at the same time keeping their overall costs at a minimum. A service provider can lose a customer if the services they provide do not meet the customer's expectations. At the same time, the service provider must operate efficiently in order to get maximum profit. Health care facilities face similar problems. If a patient is required to wait for a long of time before receiving service, then the health care provider will eventually lose that patient to another medical provider.

Overcrowding and congestion of patients is a common challenge in many hospitals in poor countries. The most affected health service facility in most hospitals is the Intensive care Unit (ICU). Only a few government hospitals in Kenya have the ICU facility and service and the cost of providing this service is very expensive. Mostly, critically wounded patients and those who undergo major surgeries require the service and this happens at any time during the day or night. A critically sick person who requires life supporting machines service may not wait for the service because the condition gets worse in every second which may lead to the death of the patient. This will be a big loss to the family, country and the hospital. A solution needs to be found that can reduce the risk associated with having to wait for service in an ICU facility. A very practical example is of the accident patient who had to suffer for 18 hours waiting for ICU service at Kenyatta National Hospital. The end result was loss of life. Moi Teaching and Referral Hospital is the only public health facility serving the western part of Kenya and Rift Valley. The facility has only six ICU beds that are required to serve a third of the Kenyan population.

The researcher therefore did utility analysis of ICU service, by using M/M/s queuing theory to examine the size of the queue, the cost in line waiting, the cost of service and utility factor to optimize service delivery in Moi Teaching and Referral Hospital (MTRH) in Eldoret town.

1.3 Objectives of the Study

1.3.1 General Objective of the Study

The general objective of this study was to apply a queuing theory model together with an Improved Taguchi Loss Function that describes the relationship between cost of running ICU and survival of patients to determine the optimum point which benefits both the hospital and the patients in Moi Teaching and Referral Hospital in Eldoret, Kenya.

1.3.2 Specific Objectives

The specific objectives of this study were to;

i). Apply M/M/s queuing model representing the dynamics of ICU utility system to determine average time a patient takes in the system and the percentage of facility utilization.

ii). Find the equilibrium point between patient waiting cost and service cost to determine the optimum number of beds required in the facility to minimize overall costs.

iii). Determine the stability of the system using the M/M/s analysis of the expected total costs together with Improved Taguchi Loss Function in Moi Teaching and Referral Hospital in Eldoret, Kenya.

1.4 Research Questions of the Study

i). What is the average time a patient takes in the system and what is the percentage rate of facility utilization in the ICU as determined using M/M/s model?

ii). What is the equilibrium point between patient waiting cost and service cost to determine the optimum number of beds required in the facility?

iii). What is the stability of the system achieved using the M/M/s analysis of the expected total costs with Improved Taguchi Loss Function in Moi Teaching and Referral Hospital in Eldoret, Kenya?

1.5 Significance of the Study

The study is to provide sufficient information to medical managers who make decisions on the use of available limited resources to improve service offered to patients and at the same time reduce strain to the health facility in the provision of services.

Customer satisfaction is expected to improve after the study if managers apply these findings because they provides ways of minimizing the time that customers have to wait on the queue before being served and maximizing the utilization of the servers or resources. This will bring the equilibrium point between the service rate and arrival rate to optimize customer survival.

The findings of the study is also a solution to the congestion problems in the hospitals that have ICUs by suggesting the optimum number of service facilities required at the minimum possible cost.

The study again is of great benefit to the government since the mentioned benefits of the model to the patients and the hospital reduces mortality rate due to emergencies and life threatening diseases of the entire population.

The model also provides essential information, after considering the cost of equipping the ICU and studying the pattern of customer arrivals to both county governments and National government to make budgetary allocations on the provision of satisfactory health services.

Lastly the study is of great use to other researchers and academicians who will be interested in this field of study.

1.6 Theoretical Framework

1.6.1 Queuing Theory Characteristics

1.6.1.1 Arrival Characteristics

The source of arrivals for a service facility has three characteristics. The size of the source population, arrival pattern of the customers to the service facility and the behaviour in which the customers arrive (Houda, Taoufik and Hichem, 2008).

1.6.1.2 Size of the Calling Population

The calling population limited or unlimited. When the number of arrivals is a small percentage of all the potential arrivals, the calling population is unlimited. Students reporting to school, cars arriving at a fuelling station and customers arriving at a banking facility are examples of unlimited calling population. Most facilities have unlimited calling population (Houda *et al.*, 2008).

1.6.1.3 Pattern of Arrivals at the System

Customers arrive at a service facility randomly or in a pattern. For example, if one customer is arriving after every ten minutes, then the pattern of arrival is known. Customers can also arrive randomly. In this case, the arrival of the next customer is not known and each customer arrives independently. Poison distribution is frequently used to represent random arrivals of customers. Resing *et al.*, (2015) said that the arrival process of customers is usually assumed that the inter arrival times are independent and have a common distribution. In many practical situations customers arrive according to a Poisson stream (exponential inter arrival times). Customers may arrive one by one, or in batches. An example of batch arrivals is the customs once at the border where travel documents of bus passengers have to be checked.

1.6.1.4 Behaviour of the Arrivals

Customers who arrive at a service facility are assumed to wait for service without bulking or reneging. A customer is said to balk when the customer refuses to join a queue because of the length. On the other hand, a customer is said to have renege if the customer gets impatient while on the queue and decides to leave without getting service. This behaviour is common as we have often seen customers in the super markets leave their goods during busy days when the queue is very long without getting the service. Most queuing models assume that customers will join the queue and patiently wait for service. This is the reasons why analysis of the queue is should be done in order to improve customer satisfaction and reduce loss due to balking and reneging (Baun and Breuer, 2005).

1.6.2. Waiting Line Characteristics

The waiting line is the number of customers in the queue waiting for service. The length of the line can be unlimited or limited. A line becomes a limited queue when it is restricted. A good example is admitting students in a class that can only accommodate 40 students. The admitting person is forced to send away any other student after receiving the required 40. Unlimited queue is a line that is allowed to grow to any length. Such a case can be seen on a road where any number of vehicles can pass without any restriction. Resing *et al.*, (2015) state that a customer may be patient and willing to wait for a long time. Or customers may be impatient and leave after a while. For example, customers call their customer service line of a mobile call service provider to get assistance, the same customers will hang up when they have to wait too long before an operator is available, and they possibly try again after a while. But in our case, critically sick patients may be transferred to other facilities, others may get better and others may die as they wait for service which affects the waiting line.

1.6.2.1 Queue Discipline

This is the rule of how customers in the line are served in a service facility. Most models use First in First Out (FIFO), where the first customer to arrive will be the first to be served. A queue in front of the checkout counter of a supermarket may serve as the simplest illustration for a queuing system. There is one input stream, and one server who serve the customers in order of their appearance at the counter. This service discipline, which does not admit any preferences among users, is FIFO(Baun and Breuer, 2005). The second discipline is serving customers in random order. For example, critically sick patients are allowed to be served first. Another discipline is Last In First Served (LIFS) also known as Last In First Out (LIFO), is common when materials are piled so that the items on top are used first. Other service disciplines include; Hold on Line (HL) where an important customer takes the head of the queue immediately he arrives. Pre-emption (PR), this happens when an important customer arrives and it is served immediately and the customer under service returns to the queue. Processor Sharing (PS), all customers are served simultaneously with service rate inversely proportional to the number of customers (Abate, 1995).

1.6.3. Service Facility Characteristics

The service facility is the third part of any system that deals with queuing. Service systems are usually classified in terms of their number of channels, or number of servers, and number of phases, or number of service stops that must be made by a customer (Aronsky and Hoot, 2008).

1.6.3.1 Kendall Notation

Kendall according to Houda *et al*., (2008) developed a notation that has been widely accepted for specifying the pattern of arrivals, the service time distribution, and the number of channels in a queuing model. This notation is often seen in software for queuing model. The basic three-symbol Kendall notation is in the form: arrival distribution/service time distribution and number of service channels open. Specific letters are used to represent probability distributions. An abridged version of this convention is based on the format A/B/C/D/E/F. These letters represent the following system characteristics:

A *= represents the inter arrival-time distribution,* ***B*** *= represents the service-time distribution. [Common symbols for* ***A*** *and* ***B*** *include* ***M*** *(exponential),* ***D*** *(constant or deterministic),* ***Ek*** *(Erlang of order k), and* ***G****(arbitrary or general)].* ***C or S*** *= represents the number of parallel servers.* ***D*** *= represents the queue discipline.* ***E*** *= represents the system capacity.* ***F*** *= represents the size of the population.*

In our case ***M*** *is used to represents the inter arrival-time distribution and the service-time distribution while S is used to represent the number of servers (Houda et al., 2008)*

1.6.3.2Single-Channel, Single-Phase System

This is a system with one queue and one service facility according to Aronsky and Hoot, (2008). A good example is a supermarket with one paying point where all customers queue to make payments as illustrated in Figure 1.

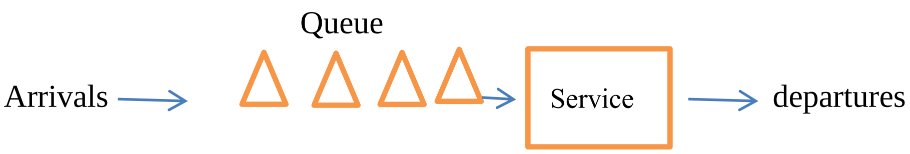

Figure 1.1Single-Channel, Single-Phase System, Source (Author)

1.6.3.3Single-Channel, Multiphase System

A food restaurant which requires you to place your order at one point, pay at a second, and pick up the food at a third service stop, becomes a multiphase system with a single channel if it has only one queue(Aronsky and Hoot, 2008).

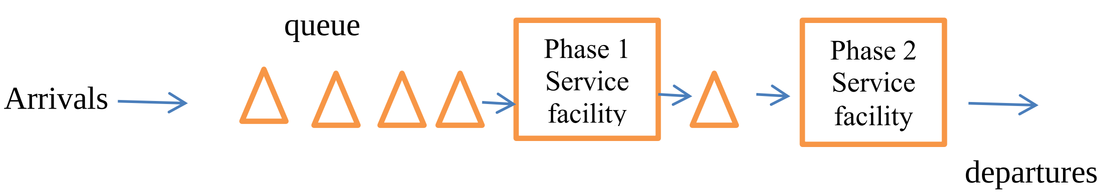

Figure1.2Single-Channel, Multi-Phase System, Source (Author)

1.6.3.4 Multichannel, Single-Phase System

This system has many service facilities where a customer is served once. Many banks today are multichannel service systems where customers form one queue in front of the tellers and the first customer in the queue get served in the next available teller. With advance in technology, these banks are now shifting to automated queueing system. Airline ticket counters is also another examble of a single channel,single phase queuing system (Aronsky and Hoot, 2008).

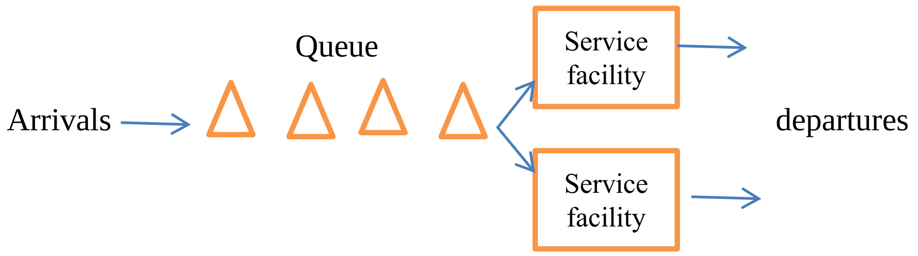

Figure 1.3Multichannel, Single-Phase System, Source (Author)

1.6.3.5 Multichannel, Multiphase System

In this system, a customer can be served in any of the many service facility and proceeds to another queue to be served in the next service facility. A good example is a dinner where guests are served from many service points and with different types of food being served (Aronsky and Hoot, 2008).

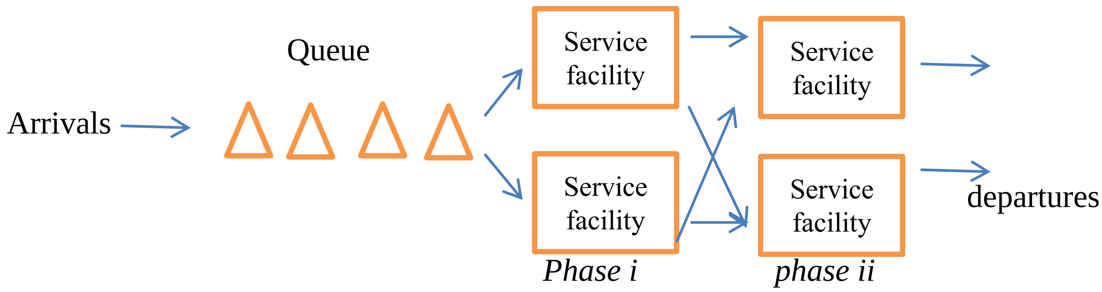

Figure 1.3Multichannel, Multiphase System, Source (Author)

1.6.4 Service Time

Service time patterns are like arrival patterns of customers. They can also be either constant or random. The time it takes to serve a customer can be fixed or random and an exponential distribution is often used. Constant service time means the amount of time taken to serve one customer is the same to time used to serve all the other customers. This is the case in services using machines such as an automatic car wash. More often, service times are randomly distributed like serving a customer in a bank or voters in a polling station. Resing *et al*., (2015) deduces that usually we assume that the service times are independent and identically distributed, and that they are independent of the inter arrival times. For example, the service times can be deterministic or exponentially distributed. It can also occur that service times are dependent of the queue length. For example, the processing rates of the machines in a production system can be increased once the number of jobs waiting to be processed becomes too large.

1.7Definition of Terms Used in the Study

A **model** is a representation of reality using mathematical concepts and language.

Queuing theory is the mathematical study of waiting lines.

Utility analysis is the evaluation of the proportion of the time that service facilities are in use.

Emergency medical services are immediate medical attention that patients may require due to an operation, accident or any other serious health condition.

Calling Population is the population of items from which arrivals at the queuing system come.

ICU bed – is a complete medical bed equipped with lifesaving machines.

Poisson distribution is a probability distribution that is often used to describe random arrivals in a queue.

Service Cost is the cost of providing a particular level of service.

Utilization Factor is the proportion of the time that service facilities are in use.

Waiting cost is the cost of having customers or objects waiting to be served.

Negative Exponential Probability Distribution is a probability distribution that is often used to describe random service times in a service system.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents literature that has been reviewed by the researcher on Queuing Theory. It shows what other researchers have done in the same area of interest to the researcher.

2.2 Queuing Theory

The study of waiting lines is a technique of quantitative analysis named as queuing theory which is widely used in many service facilities in making decisions (Fomundam and Herrmann,2007). Queuing theory is generally considered as a branch of operations research because the results are often used when making business decisions about the resources needed to provide a service.

The study requires the analysis of parameters which manage planning and selection of equipment in order to decide equipment requirement (type and optimum number), waiting time, idle time and time spent in system. Balancing the cost of providing services with the costs of customer waiting is the decision problem involved here. Use of queuing theory in healthcare is now utilized worldwide (Gillett, 2006). Research has shown that queuing theory can be useful in real-world healthcare situations, and reviews of this work have appeared. A queue in the more exact scientific sense consists of a system into which there comes a stream of users who demand some capacity of the system over a certain time interval before they leave the system again (Baun and Breuer, 2005).

A considerable body of research has shown that queuing theory can be useful in real-world healthcare situations, and some reviews of this work have appeared. Many researchers have reviewed queuing models for evaluating the impact of bed assignment policies on utilization, waiting time, and the probability of turning away patients. Nosek and Wilson (2008) reviewed the use of queuing theory in pharmacy applications with particular attention to improving customer satisfaction. There work mainly focused on the time a customer waited to be served and how to reduce the same time to a level that a customer got satisfied.

Fomundam and Herrmann, (2007) described the contributions and applications of queuing theory in the field of healthcare. They summarized a range of queuing theory results in areas of waiting time and utilization analysis, system design and appointment system. It is abundantly clear that waiting line model has come to be used in healthcare system. Originally developed for analysing the telephone traffic density, waiting line model has now found tremendous applications in almost all the service areas such as ATM, Banks, Petrol pumps queues, retail shops. Queues find further applications in airport traffic. Here, the servers are the several landing fields available for arriving airplanes, while the latter are the users of the system. Obviously, there cannot be any queue of planes waiting in the air, so that an arriving airplane finding all landing fields in use needs instead to fly an extra circle around the airport and then try again for a possibility to land. Such a manoeuver is called a retrial, and the corresponding queuing model is called a retrial queue. Since with every extra circle that a plane has to perform its gasoline is reduced more, the priority of such an aircraft to obtain a landing permission is increasing and should be higher than that of more recent airplanes with fewer retrials. Such an influence on the service schedule is called priority queuing (Baun and Breuer, 2005). A study done by Green, Kolesar, and Whitt, (2007) examined the effectiveness of a queuing model in identifying provider staffing patterns to reduce the fraction of patients who leave without being seen and their conclusion was that queuing models can be extremely useful in most effective allocation of staff.

Effective resource allocation and capacity planning are determined by patient flow because it informs the demand for health care services (Murray, 2000). Queuing theory provides exact or approximate estimation of performance measures for such systems based upon specific probability assumptions. In a hospital, these assumptions rarely hold, and so results are approximated (Cochran and Bharti, 2006).

McClain (1976) reviews research on models for evaluating the impact of bed assignment policies on utilization, waiting time, and the probability of turning away patients. Nosek and Wilson (2008) review the use of queuing theory in pharmacy applications with particular attention to improving customer satisfaction. Customer satisfaction is improved by predicting and reducing waiting times and adjusting staffing. Preater, (2002) presents a brief history of the use of queuing theory in healthcare and lists many papers that have been written on it. However, it provides no description of the applications or results. Green, (2006) presents the theory of queuing as applied in healthcare. She discusses the relationship amongst delays, utilization and the number of servers, the basic M/M/s model, its assumptions and extensions; and the applications of the theory to determine the required number of servers. For example, understanding how to model a multiple-server queue, could make it possible to determine how many servers are actually needed and at what wage in order to maximize financial efficiency. Or perhaps a queuing model could be used to study the lifespan of the bulbs in street lamps in order to better understand how frequently they need to be replaced. The applications of queuing theory extend well beyond waiting in line at a bank (Kembe *et al.*, 2012).

It may take some creative thinking, but if there is any sort of scenario where time passes before a particular event occurs, there is probably some way to develop it into a queuing model. Queues are so commonplace in society that it is highly worthwhile to study them, even if only to shave a few seconds off one's wait in the checkout line. The researcher agrees that queuing theory is of valuable use in evaluating health care facilities and will use it to solve the problem at hand. Queuing theory can be applicable in many real world situations.

2.3 Model design

Our most important objective when designing a healthcare system is reducing waiting times because long patient wait on the queue is undesirable. This then has led many researchers interested in service delivery to design a model that can determine system capacity based on desired system goals and requirements. The variables of interest that can be measured using the model are usually staffing levels, beds, or other key resources.

At the Dallas bureau, statistics shows that customer waiting time for birth and death certificates reduces by decreasing the time required to serve each customer says (Moore, 1977). The researcher first used queuing theory to calculate the service rate required to achieve a target waiting time of 15 minutes. Then this service rate is converted to the time required to serve one customer. The reduced time required to serve each customer is attained through the introduction of new equipment and more efficient processes.

Agnihothri and Taylor (1991) investigated scheduling department that handles phone calls whose intensity varies throughout the day to seek the optimal staffing at the hospital and putting into consideration known peak and non-peak periods of the day. They grouped periods that received same call intensity and determined the necessary staffing for each such intensity, so that staffing varies dynamically with call intensity.

As a result of reorganizing server capacity over time, customer waiting time reduced and complaints immediately reduced without an addition of staff. Green *et al.*, (2007) also used the same approach and named it Stationary Independent Period by Period to adjust staffing in order to reduce the percentage of patients that renege. He however, argues that congestion starts after the arrival peak, the staffing levels should lag behind the service demand levels.

2.4 Variable of Arrival Rate

Arrival rate is the number of arrivals in any given time period of the patients requiring emergency service in a health facility. In our case, the arrival is random and follows the Poisson distribution.

According to Karlin and McGregor (1988), the Poisson distribution was named after the famous French Mathematician, Simeon Denis Poisson (1781-1840) who first studied it in 1837. He applied it to results such as the probability of death in the Prussian army resulting from the kick of a horse and suicides among women and children. The Poisson process is considered the most "random" arrival process because of its assumption that the number of arrivals in any given time period, which has a Poisson distribution, is independent of the number in any other non-overlapping time period.

Rosenquist, (1987) studied how an increase in patient arrival rate affected waiting times and queue length for an emergency radiology service. A system with congestion discourages arrivals. Worthington (1991) argues that increasing service capacity which is the traditional method of attempting to reduce long queues has little effect on queue length because as soon as patients realize that waiting times would reduce, the arrival rate increases, which increases the queue again. Many healthcare systems have a variable arrival rate though some models assume a constant arrival rate. In some cases, the arrival rate may depend upon time but be independent of the system state. For example, arrivals change due to the time of day, the day of the week, or the season of the year. In healthcare, the Poisson process has been verified to be a good representation of unscheduled arrivals to various parts of the hospital including ICUs and obstetrics units.

2.5 Priority Queuing Discipline

In a grocery checkout line, any arrival is added to the end of the queue and service is not performed on it until all of the arrivals that came before it are served in the order they arrived. Although this is a very common method for queues to be handled, it is far from the only way. The method in which arrivals in a queue get processed is known as the queuing discipline Biggs (2008). This particular example outlines a First-Come-First-Serve discipline, or an FCFS discipline. Other possible disciplines include Last-Come-First-Served or LCFS, and Service In Random Order, or SIRO. While the particular discipline chosen will likely greatly affect waiting times for particular customers for instance nobody wants to arrive early at an LCFS discipline, the discipline generally doesn't affect important outcomes of the queue itself, since arrivals are constantly receiving service regardless.

According to Biggs (2008) Elective surgery waiting lists are used to manage access to public hospital elective surgery services and give priority to those in most urgent need of care. They have become an integral feature of our health system, and allow limited health resources to be allocated or 'rationed' on the basis of need. Waiting lists also provide health consumers with an indication of how long they can expect to wait for their surgery.

Siddhartan, Jones and Johnson, (1996) proposed a priority discipline for different categories of patients and then a first-in-first-out discipline for each category. They found that the priority discipline reduces the average wait time for all patients.

However, while the wait time for higher priority patients reduced, lower priority patients endured a longer average waiting time.

Taylor *et al*., (1989) modelled an emergency anaesthetic department operating with priority queuing discipline. They were interested in the probability that a patient would have to wait more than a certain amount of time to be served. Haussmann, (1970) investigated the relationship between the composition of prioritized queues and the number of nurses responding to inpatient demands. The authors found that a slight increase in the number of patients assigned to a nurse with a patient mix with more high-priority demands resulted in very large waiting times for low priority patients.

McQuarrie, (1983) showed that it is possible, when utilization is high, to minimize waiting times by giving priority to clients who require shorter service times. This rule is a form of the shortest processing time rule that is known to minimize waiting times. It is rarely found in practice due to the perceived unfairness unless that class of customers is given a dedicated server, as in a bank with a dedicated teller to customers with bulk money. Worthington (1991)analysed patient transfer from outpatient physicians to inpatient physicians. The patient was assigned one of three priority levels. Based on the priority level, there was a standard time period before which a referred patient should be scheduled to see the inpatient physician. The model assumed sufficient in-patient capacity to treat the highest priority category within. All these queuing priorities are applicable in many situations. The researcher used FIFO discipline in the study.

2.6 M/M/1 Model

This is a Single Channel Queuing model with Poisson arrivals and Exponential service time. This is the most common case of a queuing problem which involves a single-channel (single –server) waiting-line. In this model arrivals form a single-line to be served by a single server. We assume the following conditions exist in this type of system: Arrivals are served on FIFO basis, and every arrival completes service regardless of the queue length. Arrivals are independent of other following arrivals, but the average number of arrivals (arrival rate) remains the same. A Poison probability distribution is used to describe arrivals that come from unlimited calling population. The time used to serve one customer is not the same to the time used to serve the next customer, service time of each is independent of one other, but their average time is used. Negative exponential probability distribution is used to describe the random service time (Gupta,Zoreda and Kramer, 2007).

2.7 The M/M/s Queuing Model

This is a system with two or more servers (channels) available to serve arriving customers. Customers wait for service form one single line and then go on to be served in any of the available server. This model assumes that arrivals follow a Poisson Probability distribution and that service times are exponentially distributed. Service is first come, first-served and other assumptions listed for the single-channel model also apply. Waiting-line models are useful in both manufacturing and service facilities. Analysis of queues in terms of waiting-line length, average waiting time, and other factors helps us to understand service systems and provide ways of improving their performances (Gupta, 2007).

Foster *et al*., (2010) observed that Queuing models are useful in that they provide solutions to problems of waiting that are particularly relevant in health care. More generally, they illustrate the strengths of modelling in health care research and service delivery.

The Multi Server queuing model (*M/M/s)* is deduced from the Karlin and McGregor (1988) representation for the transition probabilities. This representation allows us to study the arrival of patients, the queue length, the waiting in line cost and service cost.

These then enabled us to determine the equilibrium to optimize service and reduce costs.

Kembe *et al.,* (2012) analysed the queuing characteristics at the Riverside Specialist Clinic of the Federal Medical Centre, Makurdi using a Multi-server queuing Model and determined the Waiting and service Costs with a view to determining the optimal service level. The results of the analysis showed that average queue length, waiting time of patients as well as overutilization of doctors could be reduced when the service capacity level of doctors at the Clinic is increased from ten to twelve at a minimum total costs which include waiting and service costs. The most common objectives of studies on the clinics have included the reduction of patient's time in the system (outpatient clinic), improvement on customer service, better resource utilization, and reduction of operating costs (Gorunescu, McClean and Millard, 2002). Analysis in such cases involves, in depth analysis of the patients arrival and flow, structure of the system, manpower characteristics and the scheduling system. Appropriate queuing models are then developed and applied for process modifications, appropriate staffing, scheduling or facility changes. The M/M/s model therefore is the best placed queuing model to be used in this study based on the objectives.

McClain (1976) reviewed research on models for evaluating the impact of bed assignment policies on utilization, waiting time, and the probability of turning away patients. Nosek (2008) reviewed the use of queuing theory in pharmacy application with particular attention to improving customer satisfaction. Customer satisfaction is improved by predicting and reducing waiting times and adjusting staffing. Resing *et al*., (2015) Proposes an incremental analysis approach in which the cost of an additional bed is compared with the benefits it generates .Beds are added until the increase cost equal the benefits.

Shimshak, Gropp and Burden,(1981) considered a pharmacy queuing system with pre-emptive service priority discipline where the arrival of a prescription order suspends the processing of lower priority prescriptions. Different costs are assigned to wait-times for prescriptions of different priorities.

Gupta el (2007) chose the number of messengers required to transport patients or specimens in a hospital by assigning costs to the messenger and to the time during which a request is in queue. The author also calculated the number of servers required so that a given percentage of requests do not exceed a given wait time and the average number of patients in queue do not exceed a given threshold.

2.8 Spread Sheet Simulation

Spread sheets and software tools based on queuing theory research can automate the necessary calculations. For example, Albin, Barrett, Ito and Mueller, (1990) use the QNA software, which calculates the time that patients are in a multi-node network, server utilization, the mean and variance of the number of customers at each node, the mean and variance of waiting time at each node, the mean and variance of the number of customers in the network, and the proportion of customers at each node that arrived from other nodes.

However, discrete-event simulation permits modelling the details of complex patient flows. Jacobson, Hall and Swisher, (2006) present a list of steps that must be done carefully to model each healthcare scenario successfully using simulation and warn about the slim margins of tolerable error and the effects of such errors in lost lives. Tucker, Barone, Cecere, Blabey and Rha(1999) and Kao and Tung (1981) used simulation to validate, refine or otherwise complement the results obtained by queuing theory. Albin *et al*., (1990) show how one can use queuing theory for get approximate results and then use simulation models to refine them. We will not explore simulation studies further in this work.

2.9 The Waiting in Line Cost and the Service Cost

Resing *et al*., (2015) stated that in general we do not like to wait, but reduction of the waiting time usually requires extra investments. To decide whether or not to invest, it is important to know the effect of the investment on the waiting time. So we need models and techniques to analyse such situations.

Young (1962) proposes an incremental analysis approach in which the cost of an additional bed is compared with the benefits it generates. Beds are added until the increased cost equals the benefits. Whilst much literature is devoted to the analysis of service systems with constant mean arrival and service rates. Green *et al*., (2007) state that most actual systems today are subject to time-varying demand, where arrival rates and the number of servers vary throughout the period of operation. In subsequent years and decades, research interest in healthcare modelling through queuing theory has developed and there now exist a multitude of studies. There are nine performance measures in queuing system which are queue length, loss probability, waiting times, system time, work load, age process, busy periods, idle period and departure times (Alfa, 2010). Among those performances, the waiting time is the most used measure of system performance by customers. According to Alfa (2010) the longer the waiting time, the worse is the perception of the level quality from a customer's point of view. Hence, waiting time is a determinant of customer satisfaction (Gillett, 2006). The longer their waiting time the more they will be dissatisfied.

Customer satisfaction is improved by predicting and reducing waiting times and adjusting staffing levels. The waiting in line cost and the service cost are the utility parameters that any researcher using quantitative analysis needs to analyse in order to improve service offered.

2.10 Utility Factor and Optimal Service Cost

This section is an overview of research into using queuing theory as an analytical tool to predict how particular healthcare configurations affect delay in patient service and healthcare resource utilization with the associated costs.

Singh (2006) found that the queuing theory in healthcare organizations is very beneficial. He used Queuing model to achieve a balance or trade-off between capacity and services delays & used the POM-QM Software for to demonstrate it. In his study, Ahmed (2003) found that the accident & emergency department is the dedicated area in a hospital that is organized and administered to provide a high standard of emergency care to those in community who perceived the need for or in need of acute or urgent care including hospitals admission.

Fomundam and Herrmann, (2007) summarized a range of queuing theory results in the following areas: waiting time and utilization analysis, system design, and appointment systems. Their goal was to provide sufficient information to analysts who were interested in using queuing theory to model a healthcare process and who wanted to locate the details of relevant models. An important example of such a system is an emergency department. Broyles and Cochran (2007) calculated the percentage of patients who leave an emergency department without getting help using arrival rate, service rate, utilization and capacity. From these percentages, they determine the resulting revenue loss. Therefore waiting time and utilization analysis in a queuing system aims at minimizing the time that customers have to wait and maximizing the utilization of the servers or other resources like doctors, ICU beds, and machines in order to reduce overall costs.

2.11 Minimizing Costs

Determining server capacity by minimizing the costs in a healthcare queuing system is a special case of system design. Most of the research assigns costs to patient waiting time and to each server. After modelling the system using queuing theory, minimizing costs reduces to an exercise of finding the resource allocation that costs the least or generates the most profit.

Keller and Laughhunn (1993) set out to determine the capacity with minimal costs required to serve patients at the Duke University medical centre. They find that the current capacity is good but needs to be redistributed in time to accommodate patient arrival patterns.

Young (1962) proposes an incremental analysis approach in which the cost of an additional bed is compared with the benefits it generates. Beds are added until the increased cost equals the benefits.

Shimshak *et al*., (1981) consider a pharmacy queuing system with pre-emptive service priority discipline where the arrival of a prescription order suspends the processing of lower priority prescriptions. Different costs are assigned to wait-times for prescriptions of different priorities.

Gupta *et al*., (2007) choose the number of messengers required to transport patients or specimens in a hospital by assigning costs to the messenger and to the time during which a request is in queue. In this problem, non-routine requests are superimposed on top of routine, scheduled requests. The authors also calculate the number of servers required so that a given percentage of requests do not exceed a given wait time and the average number of patients in the queue do not exceed a given threshold.

Assuming a phase-type service distribution, Gorunescu *et al*., (2002) assign costs based on a base stock inventory policy. In this pure loss model, there is a holding cost associated with an empty bed, a penalty cost associated with each patient turned away, and a profit assigned to each day a bed is occupied.

Khan and Callahan (1993) used advertisement in their model to control the demand for laboratory services. They determined the number of clients that would maximize profits for each staffing. The staffing level with maximum profits was chosen and was applied the necessary amount of advertising that would attract the desired number of clients. The model assumes that clients would leave without service if they wait above a certain amount of time.

Rosenquist (1987) chooses staffing capacity in an outpatient radiology service with a limited waiting area by minimizing cost. He suggests scheduling patients when possible and segregating patients based on expected examination duration. Such measures would reduce variability and decrease expected waiting times.

Gorunescu *et al*., (2002) use backup beds (only staffed during peak demand) to reduce the probability of patient turn-away at a marginal cost. The model assumes a phase-type service distribution.

2.12 Optimizing Customer Survival

Gorunescu *et al*., (2002) developed a queuing model for the movement of patients through a hospital department. Performance measures, such as mean bed occupancy and the probability of rejecting an arriving patient due to hospital overcrowding, are computed. These quantities enable hospital managers to determine the number of beds needed in order to keep the fraction of delays under a threshold, and also to optimize the average cost per day by balancing the costs of empty beds against those of delayed patients. This ensures that patients are served and their survival rate is increased. McManus, Long, Cooper and Litvack,(2004) presented a medical-surgical Intensive Care Unit where critically ill patients cannot be put in a queue and had to be turned away when the facility was fully occupied. This is a special case, where the queue length cannot be greater than zero, which is called a pure loss model. Green, (2006) applied queuing models to determine the number of nurses needed in a medical ward. They are relying on queuing models such as Erlang-C and loss systems, to recommend bed allocation strategies for hospital wards. Green *et al*., (2007) survey and develop time-varying queuing networks that help determine the number of physicians and nurses required in an emergency department. If the performance of service provider falls significantly below customer's expectations, they will be dissatisfied. Some costs are incurred when service level is too low and a customer becomes dissatisfied. A low level of service may incur high cost of quality loss. Whereas, a high level of service will increase investment cost to maintain the process control, improve the process, or operator training (Plante, 2000).

The main interest of these researchers was to increase patient survival in emergency departments. In recent years, however, queuing models have been developed and used in studying multi-facility interactions and their results have positively affected the management of service facilities towards optimizing customer survival.

2.13Taguchi Loss Function

The Taguchi Loss Function (TLF) was derived by Genichi Taguchi in the late 1950s in Japan. Previous quality models had argued that no cost to the organization or the consumer was incurred unless the product went beyond its upper or lower specification limits. As analysed by (Gillett, 2006) the cost of a dissatisfied customer is not negligible, they described Waiting in line is a primary source of dissatisfaction. They mentioned that a well-known queuing theories and integrating theory behind the Taguchi Loss Function, a manager can derive the costs associated with this dissatisfaction & that customer dissatisfaction is not just an issue at the upper specification limit, but rather for each moment in time beyond the targeted wait time. They illustrated by using the Taguchi Function, it can then be seen that these costs increase beyond the upper specification limit. However, by assessing these costs and then taking measures to reduce either the actual or perceived waiting times, organizations can quantitatively determine the cost-benefit relationship of improved waiting lines. Taguchi (1986) state that, the Taguchi Loss Function approach demonstrates how reducing variation reduces costs even if all outcomes meet specification.

Taguchi's concepts led to the development of the most complete definition of quality. All characteristics should have minimum, stable variation around an optimum value. With his loss function concept, Taguchi was able to demonstrate that reducing variation below specifications was the best economic alternative.

At the time of the industrial revolution when the first attempts to make interchangeable parts were underway, one breakthrough was the invention of go and no-go gauges. For instance, testing the diameter of an axle might be done with two rings of steel, one slightly smaller than the other. The larger go gauge must fit over the axle, the smaller no-go gauge must not. This breakthrough made possible the manufacture of components in locations far removed from the point of assembly and is a cornerstone of mass production (Gillett, 2006). Soon afterwards, the concept of specifications (or tolerances) was developed. It was at this point that the first definition of quality developed: All characteristics must remain within the go, no-go specifications. This is the well-known "goal post" approach to quality. Providing all outcomes stayed between the goal posts (the specifications), all was well. Often pharmaceutical organizations believe that if a result meets specification, not only does nothing more need to be said about it, but also nothing more should be said about it. Ignoring the variation in a key characteristic because it meets specification can be a terrible but all-too-common mistake. Gillett

(2006) stated that Taguchi followed and enhanced this line of thinking. He defined the cost of poor quality as the total loss incurred by society due to variation and poor quality. Taguchi was passionate about quality to the point where he claimed that the manufacturer of poor quality (with particular reference to rework and rejects) was far worse than a thief. When a thief steals $100 from a neighbour, he has gained and his neighbour has lost, but the net economic impact in the society is nil. Regardless of who holds it, the $100 will still be invested or spent on goods and services. However, if a manufacturer throws away $100 in rejects and rework, the cost of wasted resources can never be recovered by either the service provider or by society (Gillett, 2006). In addition, a process with high levels of variation in process flow will have lower throughput that would be the case if the variation in process flow was lower. The difference in throughput translates directly into unit costs. Higher variation in process flow costs money. This loss is permanent.

To begin with, the theory that design engineers, chemists, and biologists strongly dislike variation is a useful way to introduce the Taguchi Loss Function. They always prefer perfect precision and have the best value in mind. However, all scientists understand that perfection cannot be easily achieved in practical situation.(Gillett, 2006), put it that just like any other service provider, health care providers are the same in that their concerns is aimed at providing quality care to their patients while being aware of cost of offering the service. The trick is, if patients feel they have not been treated well and that they are not getting the level of care deserved, they could move out and seek out other health care providers. By minimizing the waiting time, to an acceptable duration, the risk of losing a patient is minimized. The operating cost in every service facility is also of interest to the health care provider. The number of servers available to serve patients in the facility directly translates to the amount of waiting time and operating cost.

Godfrey, (1992) proposed that this the technique used in manufacturing environments be applied to health care problems. He argued that just as the loss function technique helped improve manufactured components, health care providers can also benefit from employing the technique. He compared the two types of situations to show that this method can be applied in the health care setting.

Some researchers have applied Taguchi's methods to health care scenarios in the past using single server channels to study the factors affecting a patient's length of stay in an Emergency Department (ED). Rinderer (1996) applied Taguchi's design of experiments methodology in an emergency medical department to determine the most significant effects on loss of service in an attempt to reduce the performance measure. He significantly cited eleven factors in the study, the three factors that stood out distinctively on the response of the loss of service were found to be having a dedicated laboratory staff, having an extra physician in the ED and implementing an auto-hold policy where a patient could be held for a while as the management try to contact his or her private physician. The number of beds, physician or resources available to serve patients directly affects the amount of time a patient spends in a health care facility. This then prompts the use of Taguchi Loss Function together with queuing theory as a tool to determine these times and queue lengths to arrive at an optimum service level agreeable to both parties (Gillett, 2006).

CHAPTER THREE

MATERIALS AND METHODS

3.1 Introduction

This chapter gives an over view of the methodology that was employed in this study and the model that was used to calculate the parameters necessary to solve the problem at hand. Data for six months was requested and obtained from Moi Teaching and Referral Hospital (MTRH) and the M/M/s model was used to calculate the parameters and spreadsheet software was used to simulate the data. An improved Taguchi Loss was then used to determine the stability of the system.

3.2General Model Characteristics and Assumptions

MTRH is a level five hospital serving more than 10 counties. The neighboring health facility of the same standards is Kenyatta National Hospital in Nairobi, which implies that the calling population is infinite. Despite the presence of competing hospitals in its proximity, the provision of emergency services which require ICU facilities is solely in MTRH except for isolated cases. The following assumptions were made for the queuing system at MTRH which is in accordance with the queuing theory. They are;

i) Arrivals follow a Poisson probability distribution at an average rate of λ customers (patients) per unit of time.

ii) The queue discipline is First-Come, First-Served (FCFS) basis by any of the servers and there is no balking or reneging. There is minimal priority classification for some extremely critical arrivals but not significantly affecting the services.

iii) Service times are distributed exponentially, with an average of μ patients per unit of time.

iv) There is no limit to the number of the queue (infinite).

v) The service providers are working at their full capacity.

vi) The average arrival rate is greater than average service rate. This is necessary to create a queue.

vii) Servers here represent doctors, beds, theatre, ICU equipment and other medical personnel necessary to provide full services to the ICU patients.

viii) Service rate is independent of line length; service providers do not go faster because the line is longer.

A model satisfying the above assumptions has the capacity to capture all the parameters that involve a multi-channel server system, where clients are served in a parallel server system. The waiting customers in a queue can be fully served if they are attended by any one of the available servers. This model could apply to many qualitative analyses of different situations. Some of the physical examples that apply include, a mobile phone provider customer care or an operator help desk, where the time on hold on the phone would represent the time in queue; and the queue length would be the number of calls that the system will accept and put on hold before giving a busy signal on the caller's mobile phone or playing a recorded message asking the caller to hang up and try again later. Also in an hospital setting, the ICU admission desk, the time waiting for a bed after a request represent the time in queue and the queue length would be the number of requests waiting for service. This can also apply to a retail store, where customers wait to be served over a counter with many cashiers. With these conditions, the most appropriate model adopted for this work is the Multi-server Queuing model (M/M/s)that can capture the dynamics of an emergency medical service with respect to utility of ICU resources.

3.3 Model Flow Chart

Following the characteristics of the hospital emergency service, and the assumptions of the model, the following flow chart represents the flow of patients in the queuing system. The system is illustrated to include *s* servers, one queue and a general ward facility for recuperating patients. In this study, the patients either admitted directly to the general ward are not considered to be in the queue, and those discharged from ICU are assumed to have left the system. Also, in case a patient admitted in the general ward becomes seriously sick and require ICU services, it is assumed that the patient will join the queue for the services. Patients in the queue are not necessarily waiting in the bench, but could be admitted in the general ward as they wait for space in the ICU facility.

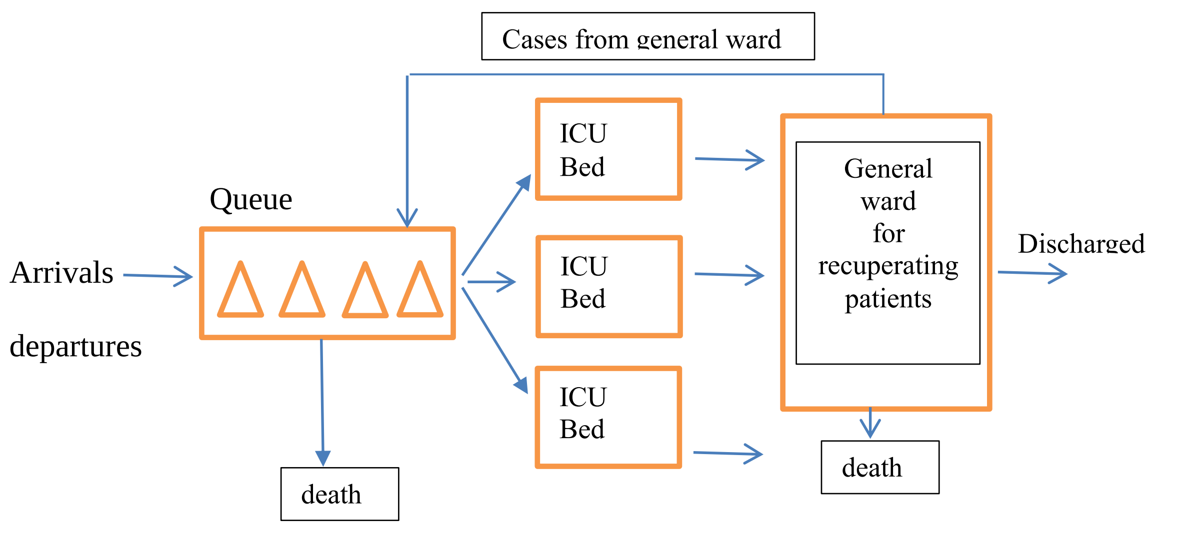

Figure 3.1Multichannel, Single-Phase System, Source (Author, 2018)

3.4 Development of M/M/1Model Equations

Before the performance measures of the service facility are worked out using M/M/s model, the assumptions and formulas of M/M/1 model are first presented that lead to the target model considering that the starting point is one service facility that will eventually give way to calculating parameters of more than one service facilities.

For this model, a systematic approach to determine important parameters of performance is presented in the service system. After calculating these numeric measures, it then becomes possible to add in cost data obtained from the model and use them to make decisions that balance desirable service levels with waiting line and service costs.

3.4.1Assumptions of the M/M/1Model

The assumptions of this model are; **a**rrivals are served on a First-In, First-Out (FIFO) basis, and every arrival waits to be served regardless of the length of the line or queue, also arrivals are independent of preceding arrivals, but the average number of arrivals (arrival rate) does not change over time. It is also assumed that arrivals are described by a Poisson probability distribution and come from an infinite or a very large population. Also service times vary from one customer to the next and are independent of one another, but their average rate is known and service times occur according to the negative exponential probability distribution. Lastly the service rate is faster than the arrival rate (Gupta, 2007).

3.4.2M/M/1 Queuing Equations

To determine the properties of this single channel, you find an expression that represents the probability of n customers in the system at time t represented by $p_n(t)$. But you shall first find the value for $p_n(t + td)$.

The probability of n customers in the system at time $t + dt$ can be determined by summing up probabilities of all the ways this event could occur. The event can occur in four mutually exclusive ways (Gupta, 2007).

Table3.1*Probability of n Customers in the System at Time t+dt*

Event	Number of units at time t	Number of arrivals in time dt	Number of services in time dt	Number of units in time $t + dt$
1	n	0	0	n
2	$n + 1$	0	1	n
3	$n - 1$	1	0	n
4	n	1	1	n

Now we compute the probability of occurrence of each of the events, noting that the probability of a service or arrival is μdtor $\lambda dt$and $(dt)^2 \rightarrow 0$

Probability of event 1 = Probability of having $n$units at time t

× Probability of no arrivals

× Probability of no services

$$= p_n(t).(1 - \lambda dt)(1 - \mu dt)$$

$$= p_n(t).[1 - \lambda dt - \mu dt + \lambda \, \mu \, (dt)^2]$$

$$= p_n(t).[1 - \lambda dt - \mu dt]$$

Similarly Probability of event 2 $= p_{n+1}(t).(1 - \lambda dt).(\mu dt)$

$$= p_{n+1}(t).(\mu dt),$$

Probability of event 3 $= p_{n-1}(t).[\lambda dt].(1 - \mu dt)$

$$= p_{n-1}(t)[\lambda dt],$$

Probability of event 4 $= p_n(t).(\lambda dt)(\mu \, . dt)$

$$= p_n(t).[\lambda \, . \; \mu \, (dt)^2] = 0$$

Note that other events are notpossible because of the small value ofdtthat causes $(dt)^2$ to approach zero, as in event 4.

Since one and only one of the above events can happen, we can obtain $p_n(t + dt)$, where $(n > 0)$ by adding probabilities of above four events

$\therefore p_n(t+dt) = p_n(t).[1-\lambda dt - \mu dt] + p_{n+1}(t).(\mu dt) + p_{n-1}(t)[\lambda dt] + 0$

$$= -(\lambda+\mu).\ p_n(t) + \mu . p_{n+1}(t) + \lambda p_{n-1}(t).$$

Taking the limit when $dt \rightarrow 0$, weget the following differential equation which givesthe relationship between p_n, $p_{n-1}(t)$, $p_{n+1}(t)$ at any time t, mean arrival rate $\lambda$and mean service rate μ;

$$\frac{d}{dt} p_n(t) = \lambda p_{n-1}(t) + \mu\, p_{n+1}(t) - (\lambda+\mu) p_n(t), \text{ where } n \rightarrow 0 \qquad (3.1)$$

After solving for $p_n(t+dt)$ where $n > 0$, it is necessary to solve for $p_n(t+dt)$ when $n = 0$.

In this case,only two mutually exclusive and exhaustive events can occur as shown below

Table 3.2 $p_n(t+dt)$ when $n = 0$.

Event	Number of units at time t	Number of arrivals in time dt	Number of services in time dt	Number of units in time $t+dt$
1	0	0	-	0
2	1	0	1	0

Probability of event 1 = Probability of having no unit at time t

× Probability of no arrivals

× Probability of no services

$$= p_0(t) \times (1-\lambda dt) \times 1$$

Probability of event 1 = Probability of having one unit at time t

× Probability of no arrivals

× Probability of one service

$$= p_1(t) \times (1-\lambda dt) \times \mu dt$$

Taking note thatif there is no unit in the system, the probability of no service would be 1, then the probability of having no unit in the system at time $t + dt$ is given by summing up the probabilities of above two events.

$$p_0(t + dt) = p_0(t) .\ (1 - \lambda dt) + p_1(t).(\mu dt).(1 - \lambda dt)$$

$$= \mu . p_1(t) - \lambda p_0(t).$$

When $dt \rightarrow 0$, the differential equation which gives the relationship between p_0 and p_1 at any time t, mean arrival rate $\lambda$and mean service rate μ;

$$\frac{d}{dt}\ [p_0(t)] = \mu p_1(t) - \lambda p_0(t), \text{ where } n = 0 \qquad (3.2)$$

Equetions (3.1) and (3.2) provide relationships involving the probability density function $p_n(t)$for all values of n but still we do not know the value of $p_n(t)$.

Assuming that the steady condition of the system is when the probability of having no customers in the system is independent of time, then

$$p_n(t) = p_n, \ \ \frac{d}{dt}[p_n(t)] = 0$$

Therefore, for a steady state system, the differential equations (3.1) and (3.2) reduce to difference equations (3.3) and (3.4) below;

$$0 = \lambda p_{n-1}, + \mu p_{n+1} - (\lambda + \mu) p_n, \text{ where } n > 0 \qquad (3.3)$$

$$0 = \mu p_1 - \lambda p_o, \qquad \text{where } n = 0. \qquad (3.4)$$

From equation (3.4), we get; $\quad p_1 = \frac{\lambda}{\mu} p_o$

Putting $n = 1$ in equation (3.3), we get

$$\lambda p_0 - (\lambda + \mu) p_1 + \mu p_2 = 0$$

Or $\quad \lambda p_0 - (\lambda + \mu)\frac{\lambda}{\mu} p_o + \mu p_2 = 0$

Or $\quad -\frac{\lambda^2}{\mu} p_o + \mu p_2 = 0$

Or $\quad p_2 = \left(\frac{\lambda}{\mu}\right)^2 p_o$

similarly, putting $n = 2$ in equation (3.3), we get

$$p_3 = \left(\frac{\lambda}{\mu}\right)^3 p_o$$

In general, $$p_n = \left(\frac{\lambda}{\mu}\right)^n p_o for\ n > 0 \tag{3.5}$$

Now the other properties of the single - channel system can be found out as follows;

Expected (average) number of customers in the system,

$$L_s = \left(1 - \frac{\lambda}{\mu}\right)\left[\frac{\lambda/\mu}{(1-\lambda/\mu)^2}\right] = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda} \tag{3.6}$$

Expected (average) number of customers waiting in the queue,

$$L_q = L_s - average\ number\ being\ served$$

$$L_q = L_s - \frac{\lambda}{\mu} = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda}{\mu}.\frac{\lambda}{\mu-\lambda} \tag{3.7}$$

Average time a customer spends in the system queue is,

$$W_S = \frac{L_S}{\lambda} = \frac{\lambda}{(\mu-\lambda)\lambda} = \frac{1}{\mu-\lambda} \tag{3.8}$$

Average waiting time of a customer in the,

$$W_q = W_S - \frac{1}{\mu}, = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu}.\frac{1}{\mu-\lambda} \tag{3.9}$$

While the facility Utilization rate is given by,

$$\rho = \frac{\lambda}{\mu} \tag{3.10}$$

3.5 The M/M/s Model Application

This is a single channel, multi-server model with patient arrival rate and service rate per hour. For this queuing model, it is assumed that the arrivals follow a Poisson probability distribution at an average rate of λ patients per unit of time. It is also assumed that they are served on a First-Come, First-Served (FCFS) basis by any of the servers (in these case ICU beds). The service times are distributed exponentially, with an average service rate of μ patients per unit time with $s$number of servers. Customers are served in order of arrival. We suppose that the occupation rate per server is smaller than one (Resing*et al.*, 2015).$\rho = \frac{\lambda}{s\mu}$

If there are n patients in the queuing system at any point in time, then the following two cases may arise: Case one is if $n < s$, then there will be no queue. However, $(s-n)$ number of servers will not be busy. Case two is if the number of customers in the system is more than or equal to the number of servers $n \geq s$ then all servers will be busy and the maximum number of customers in the queue will be $(n - s)$. *If p_o is the probability that there are no customers (patients) in the system, p_n the probability of n customers in the system, L_q expected number of customers in the queue, $L_s$expected number of customers in the system, $W_q$expected time a customer (patient) spends in the queue, W_s expected time a customer (patient) spend in the system, then;*

λdt isprobability that an arrival enters the system between time$t$ and time$t + dt$interval and $1 - \lambda dt$is probability that no arrival enters the system within interval $t, t + dt$.

μdt is the probability of one service completion between t and $t + dt$ time interval.

Using $p_{n+i}(t)$; $i = 0, 1, 2, ...$ as the transient state probability of exactly $n + i$customers in the system at time t, assuming the system started its operation at time zero and $p_{n+i}(t + dt)$; $i = 0, 1, 2, ...$ at time $t + dt$, the properties of the Multi-channel model, it's necessary to find an expression for the probability of n customers in the system at time t. This can happen in three ways, namely when $n = 0, 1 \leq n \leq s - 1$ and $n = s - 1$.

There will be three cases in this system.

$\boldsymbol{case\ 1\ (When\ n = 0):}$

Let us first find $p_o(t+dt)$. This event can only occur in two exclusive and exhaustive ways:

Table 3.3*Probability of n Customers in the System at Time t when n=0*

Event	Number of units at time t	Number of arrivals in time dt	Number of services in time dt	Number of units in time $t+dt$
1	0	0	-	0
2	1	0	1	0

$$p_o(t+dt) = p_o(t).(1-\lambda dt) + p_1(t).(1-\lambda dt).(\mu dt)$$

$$= p_0(t) - p_0(t).\lambda dt + p_1(t).(\mu dt) - p_1(t).\lambda\mu(dt)^2$$

Noting that dt^2 and $(dt)^2 \to 0$, we get,

$$= p_0(t) - p_0(t).\lambda dt + p_1(t).(\mu dt)$$

$$\frac{p_o(t+dt) - p_0(t)}{dt} = \mu p_1(t) - \lambda p_0(t)$$

Taking the limit $\quad dt \to 0, \frac{d}{dt}[p_0(t)] = \mu p_1(t) - \lambda p_0(t)$

Considering the steady state system, $0 = \mu p_1 - \lambda p_o$.

$$p_1 = \frac{\lambda}{\mu} p_o \tag{3.11}$$

$case$ 2 ($When$ $1 \leq n \leq s-1$) :

When n lies between 1 and $s-1$, all customers arriving will be immediately served and n channels out of s will be busy. Let us first find $p_n(t+dt)$. This event can occur in three exclusive and exhaustive ways.

Table3.4*Probabilities of n Customers in the System at Time t when 1 ≤ n ≤ s-1*

Event	Number of units at time t	Number of arrivals in time dt	Number of services in time dt	Number of units in time $t+dt$
1	n	0	0	N
2	n-1	1	0	n
3	n+1	0	1	n

$p_n(t+dt) =$

$p_n(t).(1-\lambda dt)(1-n\mu dt)+p_{n-1}(t).\lambda dt.[1-(n-1)\mu dt]+p_{n+1}(t).(1-\lambda dt).[(n+1)\mu dt].$

$$=p_n(t)[1-(\lambda+n\mu)dt]+\ p_{n-1}(t).\lambda dt+p_{n+1}(t).(n+1)\mu dt$$

$$\frac{p_n(t+dt)-p_n(t)}{dt} = -(\lambda+n\mu)p_n(t)+\lambda p_{n-1}(t)+(n+1)\mu.p_{n+1}(t).$$

Considering the steady state system,

$$\lambda p_{n-1}-(\lambda+n\mu)p_n+(n+1)\mu p_{n+1}=0\ for\ 1\le n\le s-1 \tag{3.12}$$

Now equation (3.11) gives $p_1 = \frac{\lambda}{\mu}p_o$

Putting $n=1$ in equation (3.12) , we get

$$\lambda p_0-(\lambda+\mu)p_1+2\mu p_2=0$$

$$\lambda p_0-(\lambda+\mu)\frac{\lambda}{\mu}p_o+2\mu p_2=0$$

$$-\frac{\lambda^2}{\mu}p_o+2\mu p_2=0$$

$$p_2=\frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 p_o=\frac{1}{2!}\left(\frac{\lambda}{\mu}\right)^2 p_o$$

similarly, putting $n=2$ in equation (3.12), we get

$$p_3=\frac{1}{3!}\left(\frac{\lambda}{\mu}\right)^3 p_o$$

in general, $$p_n=\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_o for\ 1\le n\le s-1 \tag{3.13}$$

Case 3 (when n≥ s)

When $n=s-1$, substituting it in equation (3.12), we get

$$\lambda p_{s-2}-(\lambda+s-1)\mu)p_{s-1}+s\mu p_{s+1}=0$$

$$p_s=\frac{1}{s\mu}[\lambda+(s-1)\mu]p_{s-1}-\frac{\lambda}{s\mu}p_{s-2} \tag{3.14}$$

Now from equation (3.15),

$$p_{s-1} = \frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^{s-1} p_o$$

and $$p_{s-2} = \frac{1}{(s-2)!}\left(\frac{\lambda}{\mu}\right)^{s-2} p_o$$

From equation (3.14),

$$p_s = \frac{1}{s\mu}[\lambda + (s-1)\mu]\frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^{s-1} p_o . \frac{1}{(s-2)!}\left(\frac{\lambda}{\mu}\right)^{s-2} p_o (3.17)$$

$$= \frac{\lambda}{s\mu}\frac{1}{(s-1)!}\left(\frac{\lambda}{\mu}\right)^{s-1} p_o + \frac{(s-1)\mu}{s\mu(s-1)!}\left(\frac{\lambda}{\mu}\right)^{s-1} . p_o - \frac{\lambda}{s\mu} . \frac{1}{(s-2)!}\left(\frac{\lambda}{\mu}\right)^{s-2} p_o$$

$$= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^{s} p_o + \frac{\lambda}{s\mu} . \frac{1}{(s-2)!}\left(\frac{\lambda}{\mu}\right)^{s-2} . p_o - \frac{\lambda}{s\mu} . \frac{1}{(s-2)!}\left(\frac{\lambda}{\mu}\right)^{s-2} p_o$$

$$= \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^{s} p_{o.}$$

Similarly, when $n = s + 1$, substituting in equation (3.12) and simplifying, we get

$$p_{s+1} = \frac{\lambda}{s\mu} . p_s = \frac{\lambda}{s\mu} . \frac{(\lambda/\mu)^s}{s!} p_0 = \frac{1}{s.s!}\left(\frac{\lambda}{\mu}\right)^{s+1} . p_o$$

$$p_{s+2} = \frac{1}{s^2 .s!}\left(\frac{\lambda}{\mu}\right)^{s+2} . p_o$$

In general $$p_n = \frac{1}{s^{n-s} .s!}\left(\frac{\lambda}{\mu}\right)^{n} . p_o, \text{ for n} \geq s \quad (3.15)$$

We now need to find the value of p_0 in terms of$s, \mu and \lambda$. Then the values of $p_n\ and p_0$ can be used to develop the other equations.

To find the value of p_0, we use the relation;

$$\sum_{n-0}^{\infty} p_n = 1$$

$$\sum_{n-0}^{s-1} p_n + \sum_{n=s}^{\infty} p_n = 1$$

$$\sum_{n-0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^{n} . p_0 + \sum_{n-s}^{\infty} \frac{1}{s^{n-s} .s!}\left(\frac{\lambda}{\mu}\right)^{n} . p_0 = 1$$

$$p_0 \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^{n} + \sum_{n=s}^{\infty} \frac{s^s}{s^n s!}\left(\frac{\lambda}{\mu}\right)^{n}\right] = 1$$

$$p_0 \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{s^s}{s!}\sum_{n=s}^{\infty} x \left(\frac{\lambda}{\mu}\right)^{n}\right] = 1$$

$$p_0 \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{s^s}{s!}\left\{\left(\frac{\lambda}{s\mu}\right)^{s} + \left(\frac{\lambda}{s\mu}\right)^{s+1} + \left(\frac{\lambda}{s\mu}\right)^{s+2} \ldots \infty\right\}\right]=1$$

$$p_0\left[\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}+\frac{s^s}{s!}.\left(\frac{\lambda}{s\mu}\right)^s\left\{1+\frac{\lambda}{\mu}+\left(\frac{\lambda}{s\mu}\right)^2+\cdots\infty\right\}\right]=1$$

$$p_0\left[\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}+\frac{(\lambda/\mu)^s}{s!}.\left(\frac{1}{1-\lambda/s\mu}\right)\right]=1$$

$$p_0\left[\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!}+\frac{(\lambda/\mu)^s}{s!}.\left(\frac{s\mu}{s\mu-\lambda}\right)\right]=1$$

$$p_0=\frac{1}{\frac{(\lambda/\mu)^n}{n!}+\frac{(\lambda/\mu)^s}{s!}.\frac{s\mu}{s\mu-\lambda}}$$

Therefore;

$$\boldsymbol{p_0}=\left[\left[\sum_{\mathbf{n=0}}^{\mathbf{s-1}}\frac{(\boldsymbol{\lambda}/\boldsymbol{\mu})^{\mathbf{n}}}{\mathbf{n}!}\right]+\frac{(\boldsymbol{\lambda}/\boldsymbol{\mu})^{\mathbf{s}}}{\mathbf{s}!}\left(\frac{\mathbf{s}\boldsymbol{\mu}}{\mathbf{s}\boldsymbol{\mu}-\boldsymbol{\lambda}}\right)\right]^{-1} \tag{3.16}$$

Now the other properties of the multi-channel system can be found out.

The expected (average) number of customers in the system denoted by $L_s$will be,

$$\boldsymbol{L_s}=\frac{\boldsymbol{\lambda}.\boldsymbol{\mu}\left(\frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}\right)^{\boldsymbol{s}}}{(\boldsymbol{s}-\mathbf{1})!(\boldsymbol{s}-\mathbf{2})^2}\boldsymbol{p_0}+\frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}} \tag{3.17}$$

while the expected (average) number of customers waiting in the queue L_q is,

$$\mathbf{L_q}=\frac{\boldsymbol{\lambda}.\boldsymbol{\mu}\left(\frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}\right)^{\mathbf{s}}}{(\mathbf{s}-\mathbf{1})!(\mathbf{s}\boldsymbol{\mu}-\boldsymbol{\lambda})^2}*\ \mathbf{p_0} \tag{3.18}$$

In order to check the survival of patients, the necessary parameter, is the average time a customer spends in the system defined as,

$$\boldsymbol{W_S}=\frac{\boldsymbol{L_S}}{\boldsymbol{\lambda}}=\frac{\boldsymbol{\mu}\left(\frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}\right)^{\boldsymbol{s}}}{(\boldsymbol{s}-\mathbf{1})!(\boldsymbol{s\mu}-\boldsymbol{\lambda})^2}*\ \boldsymbol{p_0}+\frac{\mathbf{1}}{\boldsymbol{\mu}} \tag{3.19}$$

Before a patient is served, the patient is expected to wait in the queue defined as,

$$\boldsymbol{W_q}=\frac{\boldsymbol{L_q}}{\boldsymbol{\lambda}}=\frac{\boldsymbol{\mu}\left(\frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}\right)^{\boldsymbol{s}}}{(\boldsymbol{s}-\mathbf{1})!(\boldsymbol{s\mu}-\boldsymbol{\lambda})^2}\ .\boldsymbol{p_0} \tag{3.20}$$

with the chances of having to wait given by the proportion defined in form of a probability as; $p(n\geq s)=\frac{\mu.\left(\frac{\lambda}{\mu}\right)^s}{(s-1)!(s\mu-\lambda)}\ .p_0$ (3.21)

The utilization factor (ρ).The fraction of time when beds are occupied

$$\rho=\frac{\lambda}{\mu s} \tag{3.22}$$

There are very slim chances that a patient arrives and finds no queue. This happens when the service rate μ is faster than the arrival rateλ. The interpretation of this in the physical situation is that the ICU is idle, thus will have a cost impact to the facility. The chances of a customer or a patient to enter the service without waiting is given by$1 - p(n \geq s)$.

The analysis of parameters used to check the minimum number of servers necessary to meet the requirements of the patients without idle servers is obtained from the average number of idle servers given by$s$.

The utilization rate of the servers is defined by $\rho = \frac{\lambda}{s\mu}$ and thus the efficiency of M/M/s model is obtained from the ratio,

$$= \frac{Average\ number\ of\ customers\ served}{total\ number\ of\ customers}$$

3.6 Calculating Costs in the Model

The cost of quality, as related to both the product and the service, is often difficult to measure. Obviously, some costs are incurred when a customer becomes dissatisfied. However because these costs are not readily quantifiable, sometimes they remain unknown but cost benefit analysis can be used to approximate these cost.

A low level of service may be inexpensive, at least in the short run but in the long run, it may incur high cost of customer dissatisfaction such as loss of future business. A high level of service will cost more to provide services and the service provider may not be able to break even. (Rising, 2015). Observed that the amount of work in the system does not depend on the order in which the customers are served. The amount of work decreases with one unit per unit of time independent of the customer being served and when a new customer arrives the amount of work is increased by the service time of the new customer. Two major costs are therefore necessary to make decision.

In order to evaluate and determine the optimum number of servers in the system, two costs must be considered in making these decisions:

(i) Service costs

(ii) Waiting time costs of customers.

The emergency medical service cost is directly incurred while providing the services. This normally includes salaries paid to employees, cost of facilities, equipment and tools used, cost of service space, waiting space and supplies. The second entails the cost associated with the customer having to wait for service including lack of patience, opportunity cost, death while waiting, increased dissatisfaction, including cost of visiting competing institution.

In this study, the costs involving provision of emergency health service include the bed and other facilities necessary in ICU, like doctor's salary, consultation fees, support staff costs, oxygen, theatre cost and even the cost of losing a patient through death.

In order to evaluate and determine the optimum number of servers in the system, two opposing costs must be considered in making these decisions: (i) Service costs (ii) Waiting time costs of customers as discussed in section 3.2 above. Economic analysis of these costs helps the management to make a trade-off between the increased costs of providing better service and the decreased waiting time costs of customers derived from providing that service.

Denote the expected service cost by,

$$E(SC) = sC_s \tag{3.23}$$

where s is the number of servers and C_S is the service cost for each server, let the expected waiting cost of the system be

$$\boldsymbol{E(WC) = \lambda W_s C_w} \tag{3.24}$$

where; λ is the arrival rate, W_s is the average time an arrival spends in the system and

C_w is the opportunity cost of waiting by customers.

Adding equation (3.23) and (3.24) yields

$$E(TC) = E(SC) + E(WC)$$

$$TC = sC_s + \lambda W_s C_w \qquad (3.25)$$

Then the results of this model were run using the excel calculator software.

The expected total cost of the queuing model with $(1, 2, 3, \dots s)$ servers will be calculated and tabulated. Later, the results will be plotted on a graph to get the equilibrium point of optimum service versus costs as shown below.

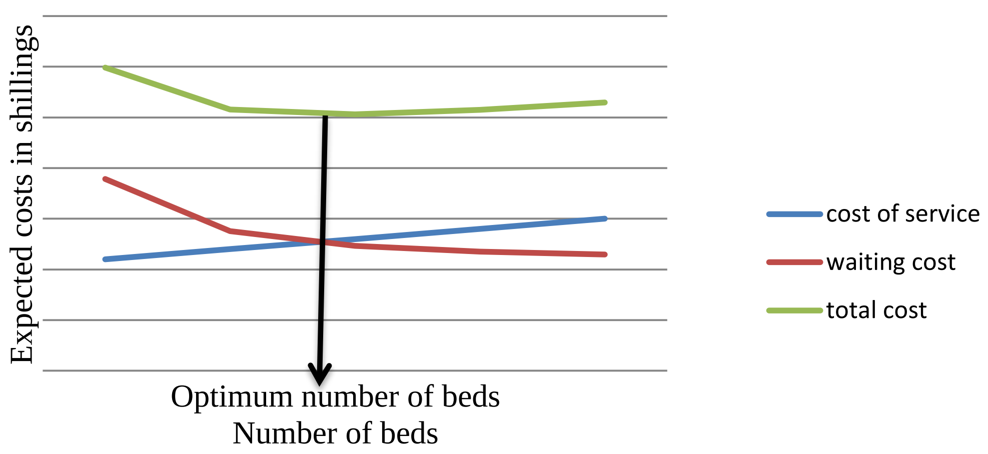

Figure 3.2 Optimum Number of Servers

3.7 Loss Function for Waiting Lines

Following the principles of formulating Taguchi loss function (Ross & Gillet, 2000), there is no cost to the service providing organization or the consumer was incurred unless the product or service went beyond its upper or lower specification limits. (USL or LSL). As expected, customers incur costs when the services provided are not meeting the expected limits, that is, the services are either too low to meet the required expectations, or too high that the consumer is not able to meet the cost. The traditional quality loss function was a square function illustrated in Figure 3.2 below.

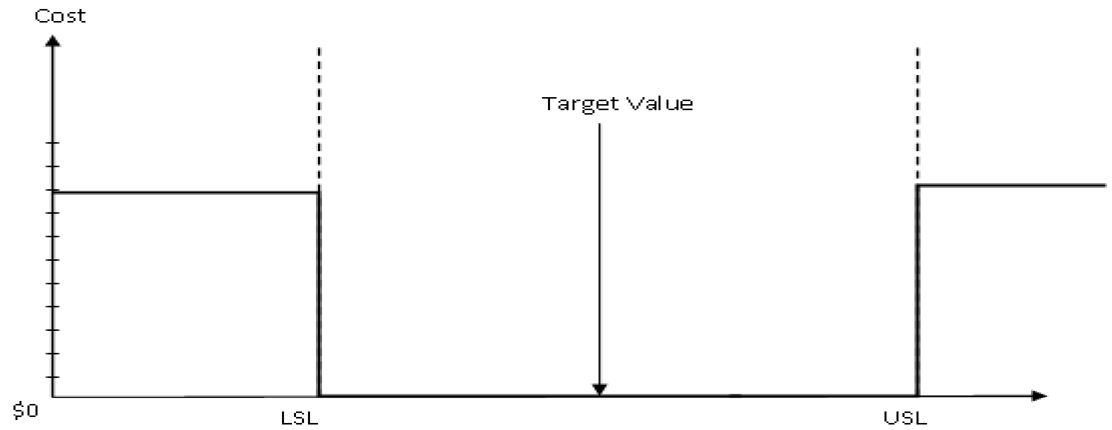

Figure 3.3The Traditional Quality Loss Function. Source (Gillett, 2006)

In this function, the customers are equally satisfied, and therefore do not incur any loss, as long as the quality of services meets the specifications between LSL and USL.

This is not realistic, and thus, an improved Taguchi loss function shown in Figure 3.3 was formulated using a quadratic function.

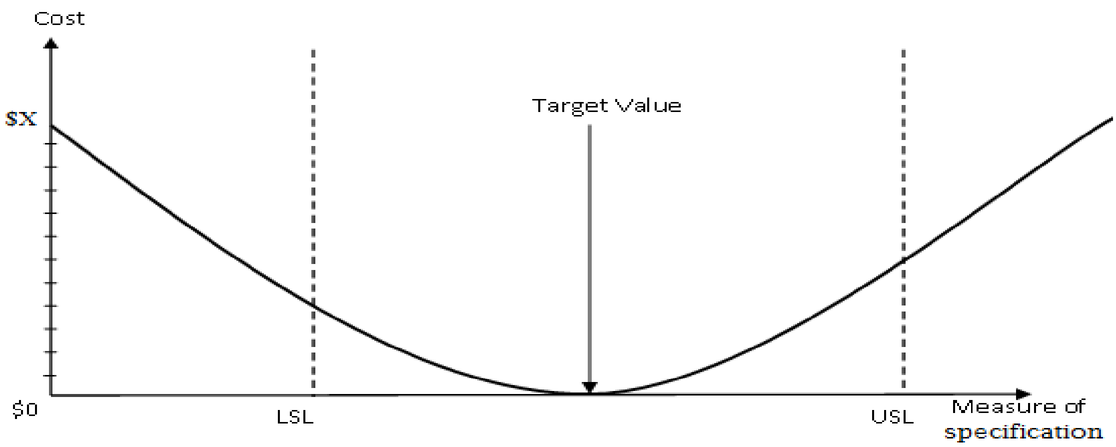

Figure 3.4Taguchi Loss Function. Source (Gillett, 2006)

The Taguchi Loss Function takes a different perspective on when the costs of poor quality are incurred. Taguchi theorized that rather than incur costs beginning at two finite points that are +/- a specific level of tolerance from the target value (or specification nominal value), costs are actually incurred as soon as the value moves from its target value. In addition, rather than continue at a constant rate, these costs are incurred at the square of the deviation from the target value, and therefore continue to increase the farther the specification deviates from the targeted value. The only point in the model at which no loss is incurred is at the actual targeted value. In contrast with traditional models, the Taguchi Loss Function is represented in Figure 3.3.

3.8 Tolerance Cost

The upper specification limit and the lower specification limits can alternatively be defined by how much a client is willing to spend for a medical service without any duress or influence. That is, drawing an horizontal line in Figure 3.4 of this minimum cost a client is willing to spend, will intersect with the cost function in two points, LSL on the left and USL on the right hand side.

However, it is obvious that people have different preferences or tastes or tolerance to unsatisfactory services. Also, due to different lifestyle and social status, the cost of waiting differs. The waiting cost is inversely proportional to the individual level of tolerance. The less the tolerance, the higher the cost of waiting. This is graphically illustrated in Fig3.5 below.

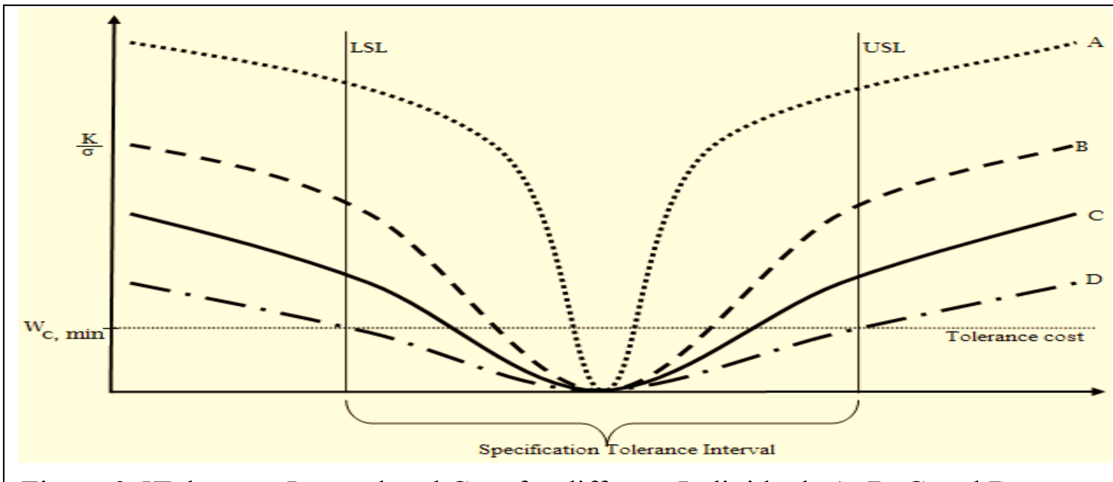

Figure 3.5Tolerance Interval and Cost for different Individuals A, B, C and D, Source (Gillett, 2006)

The tolerance parameter $\boldsymbol{\sigma}$ determines the spread of the cost function and the peak of the sides or the height where it intersects with USL and LSL lines.

3.9Determining the Stability of the System

By combining the Taguchi Loss Function with the appropriate queuing equations of calculating costs, we are able to calculate the cost of customer dissatisfaction associated solely with the time spent waiting for service and facility idle time cost and determine the Lower Specification Limit and Upper Specification Limit to define the stability of the system. Note that only the positive side of the Taguchi loss function is used for waiting time, since waiting time is only one-sided because a negative wait time is impossible, but to have a full graph, the other side of Taguchi Loss Function represents the facility idle time costs. Two derivations are provided, one using cost of time in line, and the other using idle time cost of the system. In some cases, the customer is only concerned with the time in line. For example, at an amusement park, the time in line is the primary concern. Most customers would prefer that the ride last longer, which would make the time in system longer. In other situations, the customer's concern is getting through the system as fast as possible. When your car is in the garage, you are primarily concerned with getting it back. Therefore, time in system would be the preferred measure.

The Taguchi loss function is a quadratic function which can hit the zero line on both sides of specification limits. It also has a uniform gradient for various values of quality tolerance.

3.10 Waiting Time in Queue

The probability distribution function of time in the line is defined as;

$$f(T) = \frac{1}{\sqrt{2\pi}\sigma}\left(1 - e^{-\frac{1}{2}\left\{\frac{L-T}{\sigma}\right\}^2}\right)$$

where $\rho = \frac{\lambda}{\mu}$ is the utilization factor and $(1-\rho)$ is the probability of no waiting time. Here, we require that $\mu > \lambda$. Using this probability density function with our cost loss function $f(T)$, we obtain expected cost per customer as;

$$C_q = \lambda W_s C_w f(T)$$

The total cost therefore will be

$$TC = sC_s + \lambda W_s C_w f(T)$$

3.11 Waiting Time and Idle Time Costs

In order to solve this queuing problem, the facility needs to be operated so that an optimum balance can be obtained between the customers waiting time cost and the servers idle time cost. The cost of servers' idle time in this case is the payment to be made to the servers for the period for which they are idle. While the cost of waiting in line is the loss of business to the customers during waiting or loss of customer who decides never to come again because of the length of the queue.

Waiting time losses can be reduced by increasing investment on facilities but will directly increase the cost of providing service and some servers may incur idle time costs. Its desirable then, to obtain the minimum sum of these two costs and this can be obtained by planning for the flow of customers into the facility and providing proper number of servers. If both variables are well controlled, the optimum balance of costs can be obtained.

3.12 Relationship between Level of Service and Waiting Time Costs.

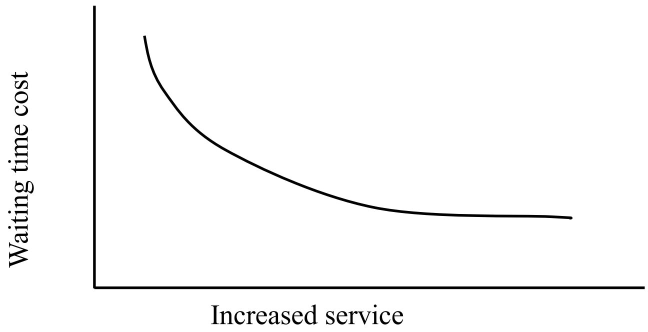

Figure 3.6 Relationship between Level of Service and Waiting Time Costs.

3.13Relationship between Level of Service and Cost of Providing Service

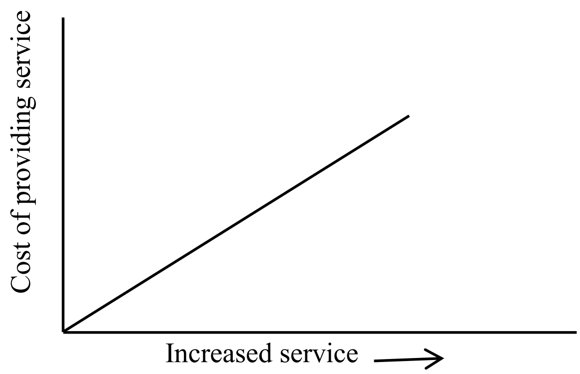

Figure 3.7*Relationship between Level of Service and Cost of Providing Service*

In this study, an infinite function is formulated which a minimum has cost as an asymptote. The cost is a function of standard deviation and the target value of expected minimum total cost. It also has a normal distribution of probabilities in a target value on the interval $-\infty \leq T \leq \infty$. The modified function is defined as;

Let C_w be the cost of rejection at the specification limit L and let T be the mean target specification value with a standard deviation of σ. Then the cost of rejection should satisfy the condition;

$$C_w = \begin{cases} Kf(\sigma, T), & L \neq T, \quad -\infty \leq L \leq \infty \\ 0, & L = T \end{cases}$$

where K is a constant of proportionality denoting the maximum cost obtained at the limit.

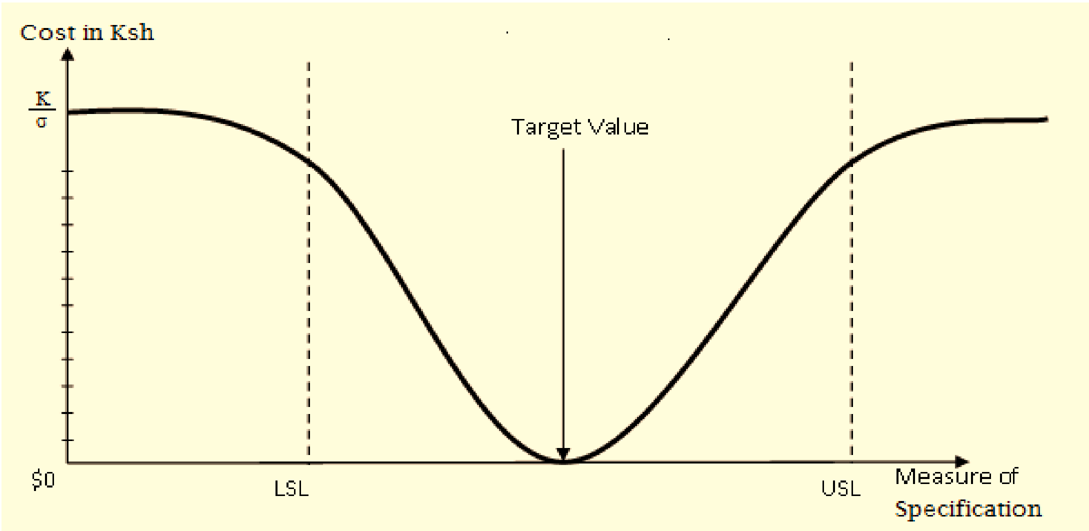

Figure 3.8Improved Taguchi Loss Function, Source (Gillett, 2006)

3.14 Estimating Waiting Cost in Relation to Tolerance Cost

The cost of waiting for an individual patient is estimated to be equal to the cost of losing a customer to other competing facilities due to congestion and dissatisfaction and the cost of patient losing life. This is because the emergency service is about life and death. In this model, it is assumed that the more the patient waits in line the more the risk of losing life. Therefore, the cost of waiting that is, equal to the average earnings of a middle class individual patient multiplied by time of waiting and evaluated per hour. This leads to a waiting cost of C_q per unit time. The total expected costs are computed using the M/M/s model and a tolerance value assumed for the patients and the facility utilization. The vertical height from the baseline (at the baseline which is the target value, there is no loss) to the loss function curve described how the amount of loss increased as results move further away from the optimum value, until eventually, complete loss occurs. Therefore the estimate of the individual cost of a patient is directly proportional to the tolerance range of the individual and the cost of losing the customer to other facilities or death. However, it is the concept of tolerance that is critical. If we understand that this variation always adds to costs.

CHAPTER FOUR

ANALYTIC RESULTS

4.1 Introduction

In this chapter, simulation of queuing costs is done in order to determine the optimum cost of the facility. The results will inform the management on the minimum number of servers required in order to reduce the waiting costs and at the same time provide service at a minimum service cost

4.2 Data Analysis

The following data was obtained from MTRH showing the bed occupancy, or number of servers and the service and arrival rates of the patients to the ICU.

Number of ICU beds	n	=6
Arrival rate of patients	λ	=5
Service rate per server	μ	=0.5
Average waiting cost	C_w	= Ksh 450
Average service cost	C_s	= Ksh 400
Total system cost	TC	= Ksh 850

4.2.1 Calculating the Dynamics of an ICU System to Determine Average Time of a Patient and System Utilization.

Computing for 8beds, 9 beds and 10 beds is impossible. This was because the arrival rate of patients was greater than the combined service rate. The problem cannot therefore be solved in that given situation. This means the queue will keep on increasing and patients who continue arriving will have to wait for long and some may not get to be served. The facility in this case cannot handle the incoming traffic.

For example, if we try to compute the probability that there are no patients in the system using equation 3.36 with 10 ICU beds, the results is as follows;

$$p_0 = \left[\left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!}\right] + \frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu - \lambda}\right)\right]^{-1}$$

$$p_0 = \left[\left[\sum_{0}^{9} \frac{(5/0.5)^n}{n!}\right] + \frac{(5/0.5)^{10}}{10!}\left(\frac{10 \times 0.5}{10 \times 0.5 - 5}\right)\right]^{-1}$$

$$p_0 = \left[\frac{10^9}{9!} + \frac{10^0}{0!} + \frac{10^{10}}{10!} \times \left(\frac{5}{5-5}\right)\right]^{-1}$$

Clearly 5-5=0, and 5/0 is infinity, meaning the facility cannot support the incoming traffic.

We therefore start calculating the parameters from 11 beds as follows,

Probability of no patients in the system

Let s =11,$\lambda = 5$ and $\mu = 0.5$

$$\mathrm{p}_0 = \left[\left[\sum_{0}^{10} \frac{(5/0.5)^{\mathrm{n}}}{\mathrm{n}!}\right] + \frac{(5/0.5)^{11}}{11!}\left(\frac{11 \times 0.5}{11 \times 0.5 - 5}\right)\right]^{-1}$$

$$\mathrm{p}_0 = 0.0000247$$

$$\mathrm{Lq} = \left[\frac{1}{(\mathrm{s}-1)!}\left(\frac{\lambda}{\mu}\right)\frac{\mu\lambda}{(\mu \mathrm{s} - \lambda)^2}\right]\mathrm{Po}$$

$$=\left[\frac{1}{(11-1)!}\left(\frac{5}{0.5}\right)\frac{0.5 \times 5}{(0.5 \times 11 - 5)^2}\right] 0.0000247$$

= 6.821

$$\mathrm{L_s} = \frac{\lambda.\mu\left(\frac{\lambda}{\mu}\right)^{\mathrm{s}}}{(\mathrm{s}-1)!(\mathrm{s}\mu-\lambda)^2}\mathrm{p}_0 + \frac{\lambda}{\mu} \quad \text{or} \quad \mathrm{L_s} = \mathrm{Lq} + \frac{\lambda}{\mu}$$

$$=\frac{5 \text{ X } 0.5\left(\frac{5}{0.5}\right)^{11}}{(11-1)!(11 \times 0.5-5)^2} 0.0000247 + \frac{5}{0.5}$$

=16.82

$$\mathrm{W_S} = \frac{\mathrm{L_S}}{\lambda}, \qquad = \frac{16.82}{5} = 3.364$$

$\mathrm{W_q} = \frac{\mathrm{L_q}}{\lambda}$, $= \frac{6.821}{5} = 1.3642$

with the chances of having to wait given by the proportion defined in form of a probability as; $p(n \geq s) = \frac{\mu.\left(\frac{\lambda}{\mu}\right)^{s}}{(s-1)!(s\mu-\lambda)} \cdot p_0$

$$= \frac{0.5.\left(\frac{5}{0.5}\right)^{11}}{(11-1)!(11 \times 0.5-5)} * 0.0000247 \quad =0.681$$

Utilization factor (ρ), representing the time the beds are occupied;

$$\rho = \frac{\lambda}{\mu s}, \quad = \frac{5}{0.5 \times 11} \quad =0.9090909$$

Developing excel calculator using the above model equation and using to run the data.

Calculating performance of 11 beds using the excel calculator was as follows;

Table 4.1*Performance of 11 Beds using the Excel Calculator*

Parameter	Value	Unit
Arrival Rate (lambda)	5	customers/hour
Service Rate per Server (mμ)	0.5	customers/hour
Number of Servers	11	servers
Average time between arrivals	0.2	hour
average service time per server	2	hour
combined service rate (s×mμ)	5.5	customers/hour
Rho (average server utilization)	0.9090909	
Po (9) Probability the system is empty)	0.00002	
L (average number in the system	16.821182	customers
Lq (average number waiting in the queue)	6.821182	customers
W (average time in the system)	3.3642364	hour
Wq (average time in the queue)	1.3642364	hour

From the table indications, under these model conditions of 11 beds, it is clear that the system is able to handle the traffic quite well, and is utilizing 90% of capacity. Traffic intensity shows that the arrival rate is 5 customers per hour and the combined service rate of 5.5 customers per hour. On average, at any given time there will be about 17 customers in the system and about 7 customers waiting in the queue. Although just over 59% of customers will have to wait in the queue, wait times are relatively brief at about 1.3 hours. Though the probability of 0.0002 chances that at any given moment the system will be full and someone will balk, or refuse to wait in the queue, and thus not enter the system at all.

4.2.2Performance Measures of the System

The results of the model in five scenarios generated using the excel calculator were as shown in the table below;

Table 4.2*Performance Measures of the Model in Five Scenarios*

No. of beds	λ	μ	p_0	ρ	L_s	Lq	W_S	W_q	Pw
11	5	0.5	0.000025	90.9	16.82	6.821	3.364	1.364	0.682
12	5	0.5	0.000036	83.3	12.247	2.247	2.249	0.449	0.449
13	5	0.5	0.000041	76.9	10.951	0.951	2.190	0.190	0.285
14	5	0.5	0.000043	71.4	10.435	0.435	2.087	0.087	0.174
15	5	0.5	0.000044	66.7	10.204	0.024	2.041	0.041	0.102

The results on the table clearly show that, all the parameters worked change when the number of beds change. The server utilization ρ drops from 100% with 6 beds to 66.7% with 15 beds. The expected average number of customers (Ls) in the system is approximately 17 customers with 11 beds and 11 customers with 15 beds. The average number of customers waiting (Lq) is approximately 7 customers with 11 beds and drops to around one customer waiting with 15 beds. The average time a customer spends in the system (W_S) is 3.4 hours with 11 beds and drops to 2 hours with 15 beds. The average waiting time of a customer on the queue is 1.4 hours with 11 beds which also drops to no waiting with 15 beds.

4.2.3 Probability of No Patient in the System

The figure below displays the probabilities of no patient in the system in five scenarios of 11,12,13,14 and 15 beds.

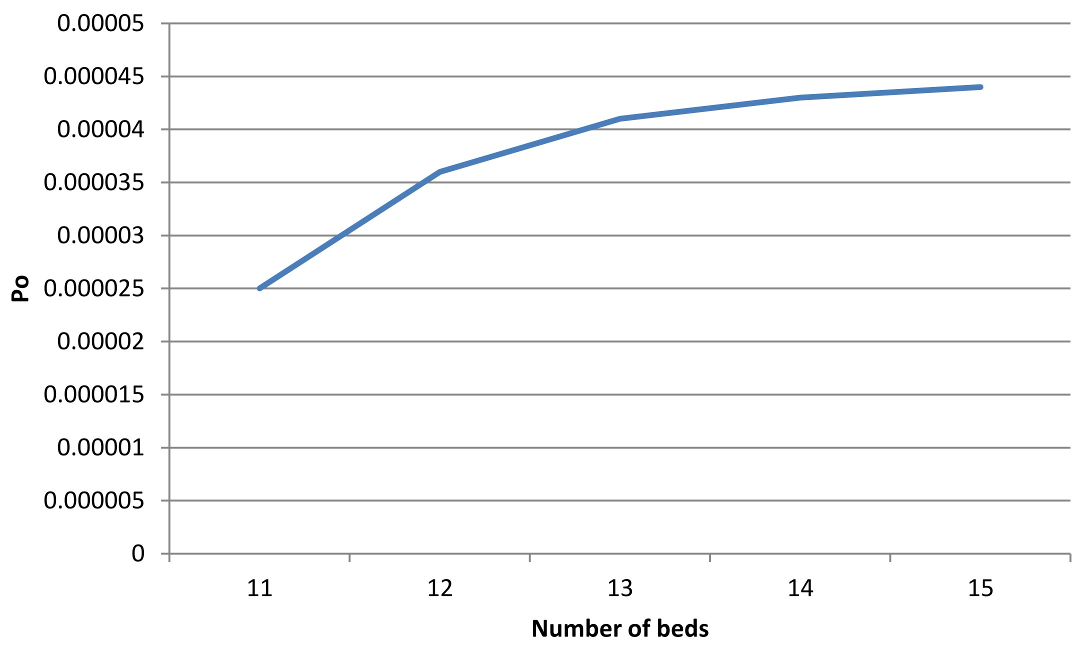

Figure 4.1 Probability of No Patient in the System

The probability of having no patient is the system result analysis is almost zero and the probability increases as the number of beds is increased.

4.2.4 Average Number of Patients in the Queue

The figure below shows the comparison of average number of patients waiting in line (queue) against number of beds in five scenarios.

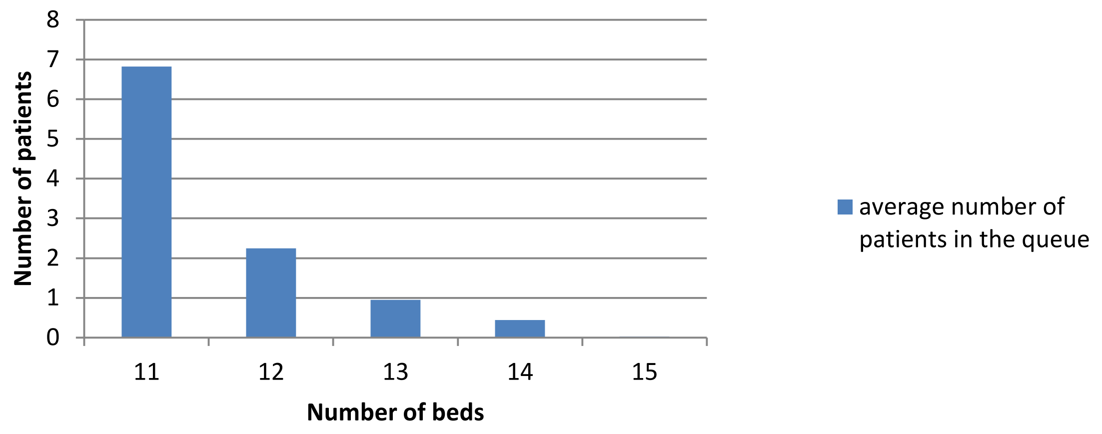

Figure 4.2 Average Number of Patients in the Queue in Five Scenarios

From the findings, we can see clearly that as you increase the number of beds, the average number of patients waiting in the line reduce. With eleven beds the number of patients is seven and with 14 and 15 beds, the number reduces to an average of one patient waiting to be served in the line. This means if we further increase the number of beds beyond 15 beds, there will be no patient waiting and that means loss to the facility due to idle servers.

4.2.4 Comparing Number of Patients in the System against Number of Beds

The figure below displays the results of the number of patients in the system per hour against number of beds in five scenarios.

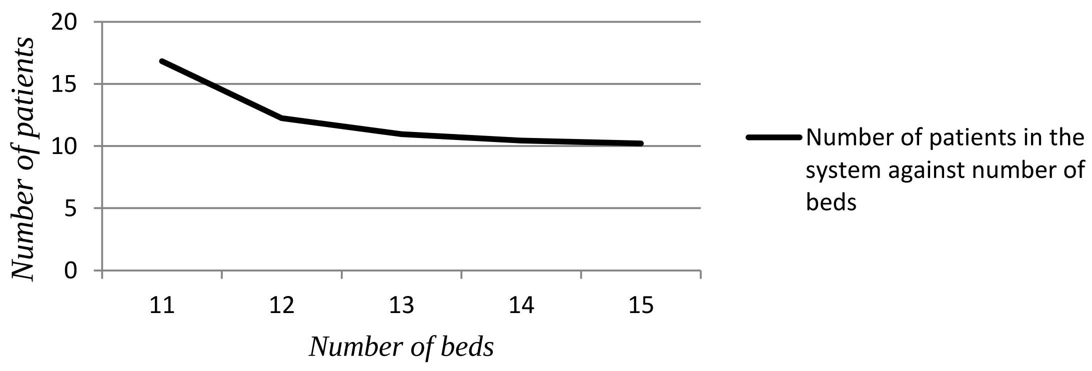

Figure 4.3 Number of Patients in the System against Number of Beds

From the graph the average number of patients in the system per hour is 17 with 11 beds, 13 with 12 beds, 11 with 13 beds, 11 with 14 beds and 11 with 15 beds. The average number in the system becomes less than the number of beds immediately the number of beds is more than 13.

4.2.5 Comparing Server Utilization against Number of Beds

The figure below represents server utilization in five scenarios compared with the number of beds in each scenario.

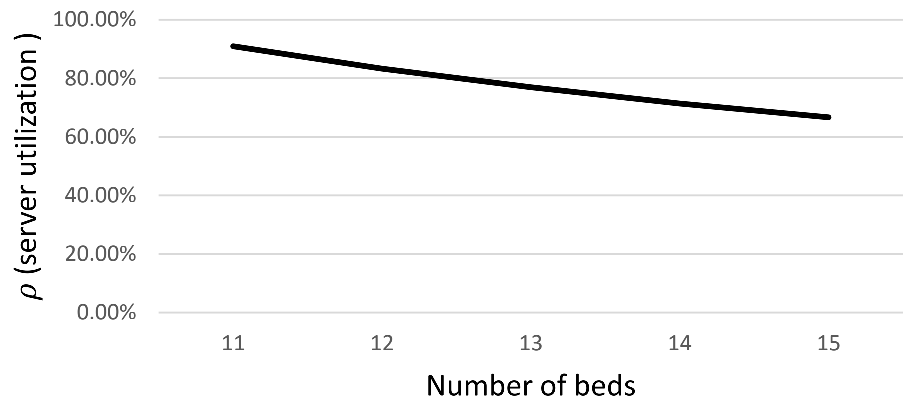

Figure 4.4 Server Utilization against Number of Beds

From the graph, the server utilization of between 66.7% to 90.9% is good to the hospital since that indicates there will be minimal idle time of the servers. However, less than 70% server utilization means increased idle time of the servers which will increase the service cost.

4.2.6 Comparing Waiting Time Against Server Utilization

The figure below shows the average waiting time of a patient per hour against server utilization in five scenarios.

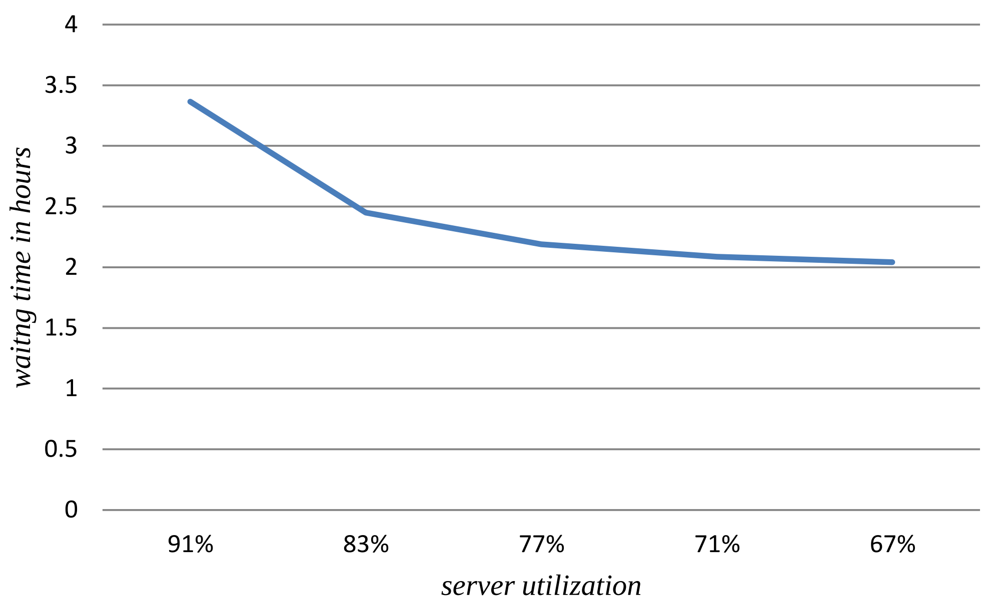

Figure 4.5 Waiting Time against Server Utilization

The graph showing comparison between the length of time a customer has to wait with server utilization indicates that the more the time a customer has to wait the less the server utilization. Here an optimum point to be identified by the planners to reduce the waiting time while maintaining good server utilization.

4.2.7 Comparing Average Patient Time in the System against Number of Beds

The figure below displays the analysis of the average time a patient spends in the system compared with the number of beds in five scenarios.

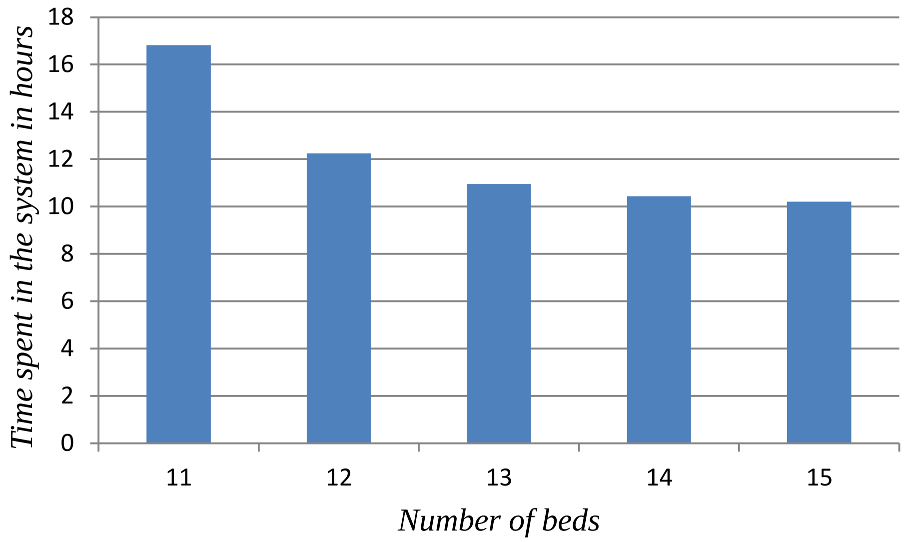

Figure 4.6 Average Patient Time in the System against Number of Beds

From the graph, we can see that the amount of time a customer has to spent in the system reduces as the numbers of servers are increased. The reduced average time is good for the customer but may increase the operating costs of the facility.

4.2.8Determining the Equilibrium Point and the Optimum Number of Beds.

Though, it is very hard to determine the cost of waiting for service, because the patient is not the only person waiting but with some other relatives waiting also, we assume that the only person incurring cost is the patient. The average cost estimate per hour includes the cost of a patient loosing life while in the system.

The actual estimate of cost of service was also a hard task to determine. MTRH is a public hospital and they use approved government rates because they receive grants to ease the load of the patients. In our case, the actual estimates of average cost of service was considered, which included the cost of the bed and its equipment's, the specialists manning the beds and the general cost of running the facility.

4.4.1Working out the Costs

The general cost estimates were;

Cost of waiting for service per our = ksh 450

Cost of offering the service per hour = ksh 400

The expected service cost is

$$E(SC) = sC_s$$

$$C_s = ksh\ 400$$

$$E(SC) = sC_s = 11 \times 400 \qquad =4400$$

waiting cost of the system is;

$$E(WC) = \lambda W_s C_w$$

$$=5 \times 3.364 \times 450 \qquad =7569.53$$

Expected Total Costs $E\ (TC)\ =\ E\ (SC)\ +\ E\ (WC)$

$$E\ (TC)\ =\ SCS\ +\ (\lambda Ws)\ Cw \qquad =4400+7569.53=11969.53$$

An excel calculator was again developed to compute the expected cost

Computed costs were as follows;

Table 4.3 *Average Patient Time in the System against Number of Beds*

No.of beds	λ	μ	W_s	C_s	C_w	$E\ (SC)$	$E\ (WC)$	$E\ (TC)$
11	5	0.5	3.36424	400	450	4400	7569.53	11969.53
12	5	0.5	2.44939	400	450	4800	5511.12	10311.12
13	5	0.5	2.19018	400	450	5200	4927.91	10127.91
14	5	0.5	2.08707	400	450	5600	4695.90	10295.90
15	5	0.5	2.04082	400	450	6000	4591.84	10591.84

The results on the table shows that as the number of servers are increased and the arrival rates and service rates remain constant, the expected service cost increases from ksh

2400 with six beds to ksh 6000 with 15 beds. The expected waiting cost on the other hand reduces from ksh 7569.50 with 11 beds to ksh 4591.80 with 15 beds. Our interest is on the total expected costs where we can see that with 11 beds the amount is ksh. 11969.50, 12 beds is ksh. 10311.10, 13 beds is ksh. 10127.91, 14 beds is ksh. 10295.50 and lastly ksh. 10591.80

4.4.2Comparing the Expected Service Cost with the Number of Beds.

The figure below shows the analysis of the expected cost of service per patient per hour against number of beds in five scenarios of 11, 12, 13, 14, and 15 beds.

Figure 4.7 Expected Service Cost against Number of Beds

As the number of beds increases the total expected service cost rises. It can be seen that with 11 beds the service cost is ksh 22,000 and the cost rises to ksh 30,000 with 15 beds.

4.4.3Analysing the Expected Waiting Cost with the Number of Beds

The figure below displays analysis of the expected waiting cost of a patient per hour in five scenarios of 11,12,13,14 and 15 beds.

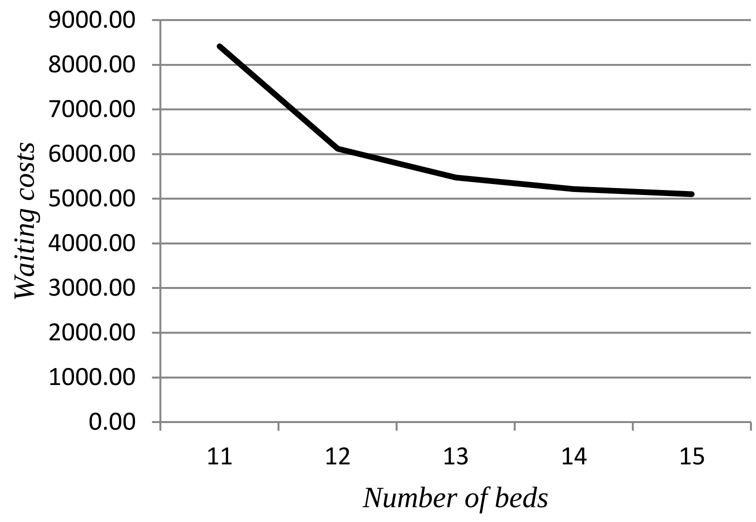

Figure 4.8 Expected Waiting Cost against Number of Beds

The expected waiting costs keep on reducing as the number of servers' increases. From the graph, we can see that the cost drops from Ksh. 8000 to Ksh. 3000 when beds are increased from 11 to 15.

4.4.4Analysing the Expected Total Cost with the Number of Beds

The figure below displays the analysis of expected total costs of the system in five scenarios.

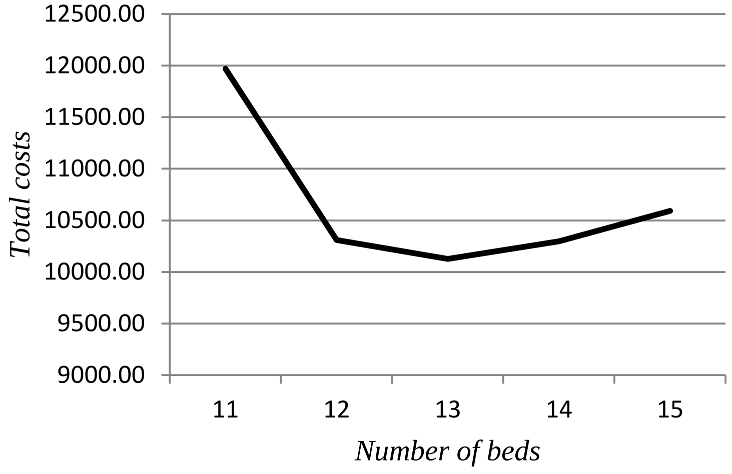

Figure 4.9 Expected Total Costs against Number of Beds

The graph of total cost indicates that, the less the number of beds the higher the total cost. It also shows that the total cost drops as the number of beds are increased to a given number but again rises as the beds keep on increasing. The optimal number of beds required that posts the lowest total cost is 13 with a total cost of approximately ksh 10,000. Increasing the number of beds to 15 is not cost effective to the hospital while reducing the number of beds to 11 will disadvantage the patient.

4.4.5Optimum Number of Beds Required

The figure below shows the analysis results of comparing the expected total costs of the system against number of beds where the optimum number of beds required is determined.

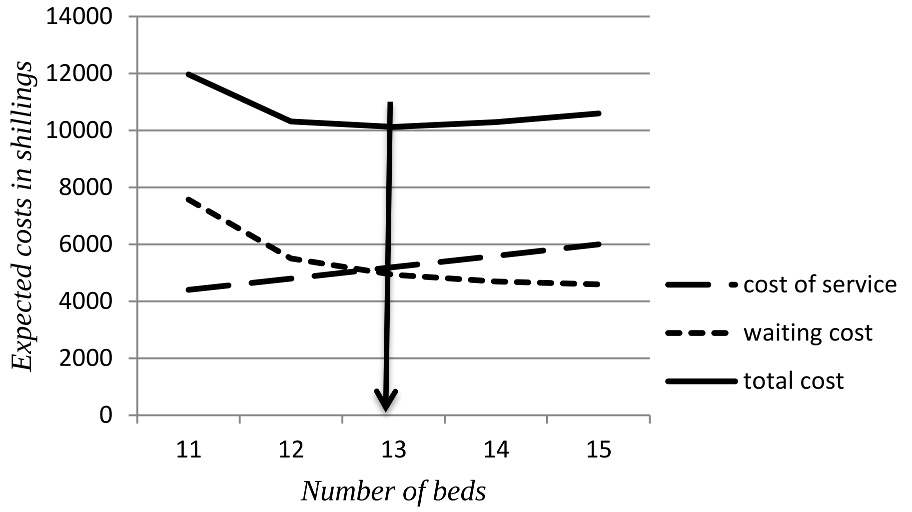

Figure 4.10 Optimum Number of Beds Required

The graph indicates that the optimum number of beds required is 13. It also shows that if we increase the beds beyond 13, the overall cost will rise up as well as the cost of service though the waiting cost will reduce. Further, the graph shows that if beds are reduced to 11, the total cost will increase as well as the cost of waiting for service though the cost of service reduces.

4.2.9 Determining the Stability of the System

Having in mind that if we only use waiting time, the positive side of the Taguchi loss function is used to determine the system stability since a negative wait time is impossible. But in this study, two derivations are provided, one using time in line, and the other using idle time of the servers. Our main interest is loss on parties, the patient and the facility. From the results of total expected cost, it is clear that costs converged to a minimum at some point. It's also evident that as you move in either direction, costs increase. We therefore use the total costs and values of normal distribution to come up with a graph to determine the stability of the system.

Assuming that each side is willing to tolerate loss of up to Ksh 300, in the total costs and considering values of total costs of the five scenarios calculated, we generate values that will be fitted into the Improved Taguchi Loss graph.

Assuming that a customer is willing to wait for a maximum of 30 minutes without complaining, and the facility does not wish to have an idle bed but wish to have 100% utilization without over straining. The results of the expected total costs were;

Table 4.4 *Expected Total Costs*

Beds	λ	ρ	μ	SC	WC	E(SC)	E(WC)	E(TC)
11	5	90.9	0.5	400	450	4400	7569.53	11969.53
12	5	83.3	0.5	400	450	4800	5511.12	10311.12
13	5	76.9	0.5	400	450	5200	4927.91	10127.91
14	5	71.4	0.5	400	450	5600	4695.90	10295.90
15	5	66.7	0.5	400	450	6000	4591.84	10591.84

Now, applying the limit of Ksh 300 tolerance of both parties, and subtracting the target value of Ksh 10127.91, we obtain costs within the limit and beyond the limit as shown below;

Table 4.5 *Cost of Rejection*

No of beds	Total Expected costs	Cost of Rejection
11	11969.53	1841.62
12	10311.12	183.21
13	10127.91	0.00
14	10295.90	167.99
15	10591.84	463.93

We then fit the costs of rejection into Taguchi Loss Function graph with Ksh 300 being the Lower Specification Limit and Upper Specification Limit with time intervals of 0.5 hours as shown

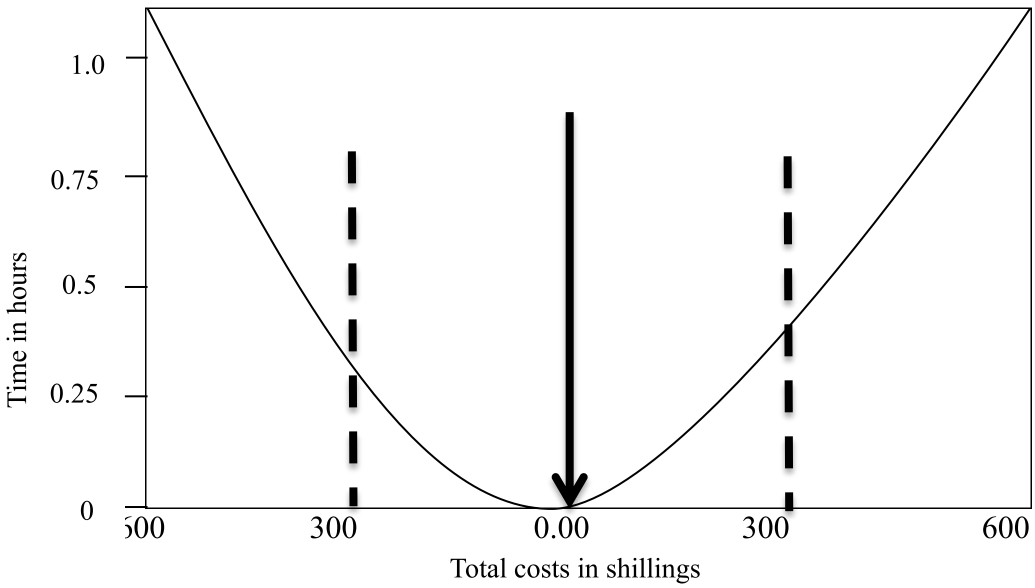

Figure 4.11 *Taguchi Loss Function*

From the graph, our target value is the least expected rejection cost, which is 0.00, equivalent to Ksh. 10,127.91 total cost in Table 4.4 above. The other rejection costs within the accepted tolerance value of Ksh 300 are Ksh 183.21 and Ksh 167.99. From our assumption that a customer is willing to loss Ksh 300, we deduce that the system is stable if the cost of rejection lies within this limit. The numbers of beds that fall within these limits are 12 to 14. Therefore the stability of the system is achieved with 12 to 14 ICU beds. Allocating more beds or fewer beds outside the limits means the service system will either be too costly to the facility or to the patient.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

This chapter presents the conclusion of the study, recommendation and further research that arises after the study to fill the gaps that were not addressed.

5.1 Summary of the Findings

The study analysed data from Moi, Teaching and Referral Hospital with an objective of addressing the queuing problem in the Intensive Care Unit that has six ICU beds. The data used was an arrival rate of five patients per hour, service rate of one patient per two hours per server and the average costs of Ksh. 400 service cost and Ksh. 450 waiting cost. Analysis of these data was done using M/M/s queuing model with an excel calculator in five scenarios starting from 11, 12, 13, 14 and 15 beds. Calculations of these data with 6, 7, 8, 9 and 10 beds was not possible with this model because the arrival rates of patients for the scenarios were more than the service rates.

5.1.1 Average Time of a Patient in the System and the Percentage of Server Utilization

The behaviour of average waiting time per patient in the queue for five scenarios was 6.821 hours with 11 beds, 2.247 hours with 12 beds, 0.951 hours with 13 beds, 0.435 hours with 14 beds and 0.024 hours with 15 beds. On the other hand, the average time a patient spends in the whole system was 16.82 hours with 11 beds, 12.247 hours with 12 beds, 10.951 hours with 13 beds, 10.435 hours with 14 beds and 10.204 hours with 15 beds. The average server utilization is 91% with 11 beds in scenario one, 84% with 12 beds in scenario two, 77% with13 beds in scenario three, 72% with 14 beds in scenario four and 67% with 15 beds in scenario five.

5.1.2 Equilibrium between Waiting Cost, Service Cost and Total Costs

The results of costs associated with patient waiting time in the system from the model for the five scenarios were, Ksh. 7569.53 with 11 beds, Ksh. 5511.12 with 12 beds, Ksh. 4927.91 with 13 beds, Ksh. 4695.90 with 14 beds and Ksh. 4591.84 with 15 beds. While the results associated service cost obtained from the model were Ksh. 4400 with 11 beds, Ksh. 4800 with 12 beds, Ksh. 5200 with 13 beds, Ksh. 5600 with 14 beds and Ksh. 6000 with 15 beds.

Again, the overall system total cost for the five scenarios generated by the model were Ksh. 11969.53 with 11 beds, Ksh. 10311.12 with 12 beds, Ksh. 10127.91 with 13 beds, Ksh. 10295.90 with 14 beds and Ksh. 10591.84 with 15 beds. Therefore, the optimum number of beds required to minimize overall cost is 13 beds.

5.1.3 Stability of the System with Improved Taguchi Loss Function Limits

Lastly, the total costs used with Improved Taguchi Loss Function shows that the target value of optimum performance is an expected rejection value of Ksh 0.00 using the tolerance value of the facility and the patient in terms of loss due to waiting time and idle time limit of Ksh 300. These Lower Specification Limit and Upper Specification Limit gave us the number of beds required for the system to be stable. And the results showed that the numbers of beds that fall within the limits are between 12 and 14. Therefore the system is stable if the bed allocation at MTRH is between 12 and 14.

5.2 Conclusion

The study has established that at the Intensive Care Unit department at Moi Teaching and Referral Hospital, the current situation of six ICU beds is inadequate and is even dangerous to the patients because some of them may never get the service in that the arrival rate of patients is greater than the service rate.

5.2.1 Average Time of a Patient in the System and the Percentage of Server Utilization

From the findings, the average waiting time of a patient per hour and the overall time spent in the system reduce as you increase the number of beds. This impacts positively to the patient and increases his survival rate because the faster you receive the service the less the risk of worsening condition.

It was also established that the server utilization remains good in the five scenarios analysed and this factor enables the hospital utilize the facility well to avoid incurring idle server costs.

5.2.2 System Costs and Optimum Number of Beds

Analysis of the costs in the system show that the cost incurred by a patient as he waits for service reduce as you increase the number of beds. This means the more the beds the better for the patient. But the cost of offering service by the facility increases as the number of beds is increased. Any extra cost with no extra income is not good for any service provider but increasing costs to offer better service will have commensurate returns, therefore the decision maker’s trade-off is good.

The study also established that the total expected cost in the system in the five scenarios is minimal with 13 beds. This indicates that optimum system performance will be achieved with 13 beds in order to address the queuing challenge by reducing waiting time and minimizing costs. Providing patients with timely access to appropriate medical care is an important element of healthcare delivery and increases patient survival.

This study also establishes that patients are generally dissatisfied with long waiting times and experience negative effects as a result which is clearly depicted by the rate at which waiting time affect waiting costs. It is further established that queuing theory and modelling is an effective tool that can be used to make decisions on staffing needs for optimal performance with regards to queuing challenges in hospitals as it was possible for us to obtain the optimum number of beds required in MTRH as 13.

5.2.3 Stability of the System

The Application of an Improved Taguchi Loss function enabled us to also conclude that the stability of the system is achieved when the bed allocation is between 12 and 14 beds. The stability is an advantage to both the facility and the patients because the tolerance level of each was accommodated which directly affects service provision and survival rates of the patients.

This study should therefore be replicated in other hospitals in Kenya and other countries in order to inform hospital administrators more on the usefulness of the application of queuing theory and modelling as a tool for improved decision making with regards to the queuing challenges that are faced by hospitals.

5.3Recommendations

In the study, it is recommended that MTRH increases their beds from 6 to 13. This is as a result of the findings that indicate minimum total cost with 13 beds. By doing this, the facility will utilize the servers well and serve the customer satisfactorily due to reduced waiting cost and waiting length. It is also recommended that the government increases funding to the facility to facilitate the acquisition of the required number of beds. Other facilities offering the same service are recommended that they do qualitative analysis of their service provision and customer satisfaction to determine the optimum service level required.

5.4Suggestions for Further Research

In this study, it is recommended that the other data necessary in an ICU be studied so that, apart from the bed, personnel working in the ICU can also be captured to determine the optimum number of doctors required and that, the same be extended to other facilities offering ICU service. To future researchers, queuing analysis is recommended as one of the most practical and effective tools for understanding and aiding decision-making in managing critical resources and should become as widely used in the healthcare community as it is in the other major service sectors. Lastly, another decision model capable of handling all situations in a service providing facility should be developed to avoid the limitation of M/M/s model that cannot work when the arrival rate is greater than the service rate.

References

Abate, J. (1995).*Numerical inversion of Laplace transforms of probability distributions. ORSA Journal on Computing*.

Ahmed, S. (2003).*Accident and Emergency Section Simulation in Hospital*. www.wseas.us/e-library/conferences/digest2003/papers/466 124.pdf

Agnihothri, S.& Taylor, P. (1991).*Staffing a centralized appointment scheduling department in Lourdes Hospital*. Interfaces 21, 1-11.

Albin, L., Barrett, J., Ito, D. & Mueller, E. (1990). *A queueing network analysis of a health center*. Queueing Systems, 7, 51-61.

Alfa, A., S. (2010). *Queueing theory for telecommunications*. Springer, New York.

Aronsky, & Hoot, R. (2008). *Review of Emergency Department Crowding: Causes, Effects, and Solutions*. Volume 52, No. 2, 16.

Bailey, J. (1954).*Queuing for medical care*. Applied Statistics, 3, 137-145.

Biggs, A. (2008). *Hospital waiting lists explained: Social Policy Section*.

Broyles, R., & Cochran, J.K. (2007).*Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression*, in Proceedings of the 2007 Industrial Engineering Research Conference.

Buan, D., Breuer, L.(2005). *An introduction to queuing theory and matrix*. Dordrecht, Netherlands: Springler.

Cochran, J., & Bharti, A. (2006). A Multi-stage Stochastic Methodology for Whole Hospital Bed Planning Under Peak Loading. *International Journal of Industrial and Systems Engineering*, (1 (1/2)), 8-35.

Fomundam S. & Herrmann J. (2007). A Survey of Queuing Theory Applications in Healthcare, University of Maryland, College Park.

Foster, E., Michael, R., & Ziya, S. (2010*). A Spoonful of Math helps the medicine Go Down : An Illustration of How Healthcare benefit from mathematical modeling and analysis*. BMC Medical Research Methodology 2010.

Gillett, F. (2006). Queuing Theory and the Taguchi Loss Function: *The Cost of Customer Dissatisfaction in Waiting Lines*. International Journal of Strategic Cost Management.

Godfrey, B. (1992). Robust Design: *A New Tool for Health Care Quality? Quality Management in Health Care,* 1(1), 55-63.

Gorunescu, F., McClean, I., & Millard, H. (2002). A queuing model for bed occupancy management and planning hospitals. Journal of the Operational Research Society.

Green, L. (2006).*Queueing analysis in healthcare, in Patient Flow*: Reducing Delayin Healthcare Delivery, Hall, R.W., ed., Springer, New York.

Green L, Kolesar P.J., and Whitt W (2007). *Coping with time-varying demand when setting sta_ng requirementsfor a service systems*. Production and Operations Management.

Gupta, I., Zoreda, J. & Kramer, N. (2007).*Hospital manpower planning by use of queueing theory. Health* Services Research, 6, 76-82.

Harrison, G., & Knottenbelt J. (2006). *A Queuing network model of patient flow*. Accident and Emergency department"

Haussmann, R.,K.,D. (1970).*Waiting time as an index quality of nursing care*. Health Services Research.

Houda, M., Taoufik, D., & Hichem, K. (2008). *Solving of Waiting lines models inthe airpor*t. 4-5.

Jacobson, S., Hall, S. & Swisher, J. (2006). *Discrete-event simulation of health care systems, in Patient Flow: Reducing Delay in Healthcare Delivery*, Hall, R.W. ed., Springer, New York, 211-252.

Kao, E.P.C. and Tung, G.G. (1981).*Bed allocation in a public health care delivery system*.Management Science.

Karlin, S. and McGregor, J.L. (1988). *Many server queuing processes with Poisson input and exponential service times*, Pacific J. Math.

Keller, F. & Laughhunn, J. (1993).*An application of queuing theory to a congestion problem in an outpatient clinic.* Decision Sciences.Matematik.

Kembe, M. M, Onah, E. S and Iorkegh, S. (2012). A Study of Waiting And Service Costs of A Multi-Server Queuing Model In A Specialist Hospital. *International Journal of Scientific & Technology Research*.

Khan R.,M., & Callahan B., B. (1993).Planning laboratory staffing with a queuing model: *European Journal of Operational Research*, Vol. 67, issue 3, 321-331.

McClain, O. (1976). *Bed Planning Using Queuing Theory Models of Hospital Occupancy: Sensitivity Analysis Inquiry*, http://www.ncbi.nlm.nih.

McManus, L., Long, C., Cooper, A. & Litvak, E. (2004).*Queuing theory accuratelymodels the need for critical care resources*. Anesthesiology.

McQuarrie, G. (1983).*Hospital utilization levels*. The application of queuing theory to a controversial medical economic problem. Minnesota Medicine.

Moore, J. (1977).*Use of queueing theory for problem solution in Dallas*, Tex., Bureau of Vital Statistics. Public Health Reports, 92, 171-175.

Murray, S. C. (2000). *Understanding the Patient Flow*. Decision Line, March, 8-10.

Nosek, R.A. and Wilson, J.P (2008*). Queuing Theory and Customer Satisfaction: A Review of Terminology, Trends and Applications to Pharmacy Practice*. Hospital Pharmacy.

Obamiro, K. (2010). Queuing theory and Patient Satisfaction*: An overview of terminology & application in Ante-Natal care unit*.http://www.upg-bulletinse.ro

Ozcan, Y. A. (2006). *Quantitative methods in health care management; Techniques and applications (First edition ed*.). Jossey-Bass Publications.

Paul Jomon Aliyas & Li Lie (2008), *Impact of Facility Damages on Hospital Capacities for Decision Support in Disaster Response Planning for an Earthquake,* pdm.medicine.wisc.edu/Volume_24/issue_4/paul.pd

Plante, R. (2000).*Allocation of Variance Reduction Targets Under The Influence of Supplier Interaction,* International Journal of Production Research, 38(12),2815-2827.

Preater, J. (2002). *Queues in health.Health Care Management Science*.

Rinderer, Z.M. (1996). *A Study of Factors Influencing ED Patients' Length of Stay at One Community Hospital*. Journal of Emergency Nursing, 22(2), 105-110.

Resing, A., Ivan, A., & Jacques, (2015). *Queueing Systems*. Eindhoven, The Netherlands: Eindhoven University of Technology.

Rosenquist, C. (1987).*Queuing analysis: a useful planning and management technique for radiology*.Journal of Medical Systems.

Schoenmeyr Tor, Dunn Peter F. and Gamarnik David (2009).*A Model forUnderstanding the Impacts of Demand and Capacity on Waiting Time to Entera Congested Recover* Room. Anesthesiology.

Shimshak, G., Gropp, D., & Burden,D. (1981).*A priority queuing model of a hospital pharmacy unit*. European Journal of Operational Research.

Siddhartan, K., Jones, J., & Johnson, A. (1996). *A priority queuing model to reduce waiting times in emergency care.* International Journal of Health Care QualityAssurance.

Singh, V. (2006). *Use of Queuing Models in Health Care Decision Analysis*.

Department of Health Policy and Management, University of Arkansas for Medical Science.

Taguchi, G. (1986). *Introduction to Quality Engineering.* Designing Quality into Products and Processes. Asian Productivity Organization.

Taylor, H., Jennings, C., Nightingale, A., Barber, B., Leivers, D., Styles, M., & Magner, J. (1989). A study of anaesthetic emergency work. Paper 1: *The method of study and introduction of queuing theory*. British Journal of Anaesthesia.

Tucker, B., Barone, E., Cecere, J., Blabey, G., & Rha, C. (1999). *Using queueing theory to determine operating room staffing needs.* Journal of Trauma, 46, 71-79.

Whitt, W. (2007). *What you should know about queueing models to set staffing requirements in service systems*. Naval Research Logistics (NRL).

Worthington, J. (1991).*Queueing Models for Hospital Waiting Lists.* The Journal of the Operation Research Society.

Young, P. (1962).*Estimating bed requirements, in A Queuing theory approach to the control of hospital inpatient census*. Technical report, John HopkinsUniversity, Baltimore

APPENDIX I

Queuing Analysis Excel Calculator

Inputs

Time unit	hour	
Ariival Rate (lambda)	5	customers/hour
Service Rate per Server (mu)	0.5	customers/hour
Number of Servers	11	servers
intermediate calculations		
Average time btw arrivals	0.2	hour
average service time per server	2	hour
combined service rate (sXmu)	5.5	customers/hour
perfomance measures		
Rho (average server utilization)	0.9090909	
Po 9Probability the system is empty)	0.00002	
L (average number in the system	16.821182	customers
Lq (average number waiting in the queue)	6.821182	customers
W (average time in the system)	3.3642364	hour
Wq (average time in the queue)	1.3642364	hour
Probability of specific nu. Of customers in the system		
Number		
probability	2.475E-05	

working calculations, mainly for Po calculations			
lambda	10		
s!	39916800		
n	(λ/μ^n	n!	sum
0		1	1
1	10	1	11
2	100	2	61
3	1000	6	227.6667
4	10000	24	644.3333
5	100000	120	1477.667
6	1000000	720	2866.556
7	10000000	5040	4850.683
8	1E+08	40320	7330.841
9	1E+09	362880	10086.57
10	1E+10	3628800	12842.31
11	1E+11	39916800	15347.52
12	1E+12	4.79E+08	17435.19
13	1E+13	6.23E+09	19041.1
14	1E+14	8.72E+10	20188.17
15	1E+15	1.31E+12	20952.89