British Journal of Mathematics & Computer Science
19(1): 1-18, 2016; Article no.BJMCS.29369
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Utility Analysis of an Emergency Medical Service Model Using Queuing Theory

T. K. Rotich[1*]

[1]Center for Teacher Education, Moi University, P.O.Box 3900 - 30100, Eldoret, Kenya.

Author's contribution

The sole author designed, analyzed and interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/BJMCS/2016/29369
Editor(s):
(1) Sergio Serrano, Department of Applied Mathematics, University of Zaragoza, Spain.
Reviewers:
(1) Wenqing Wu, Southwest University of Science and Technology, China.
(2) Hanifi Okan Isguder, Dokuz Eylul University, Turkey.
(3) Rakesh Kumar, Smvd University-Katra Shri Mata Vaishno Devi University, Katra, India.
Complete Peer review History: http://www.sciencedomain.org/review-history/16552

Original Research Article

Received: 6th September 2016
Accepted: 4th October 2016
Published: 14th October 2016

Abstract

Aims/ Objectives: To use queuing model to deter mine the optimum waiting and service cost in a hospital ICU emergency service.
Study Design: Modeling and Simulation.
Place and Duration of Study: ICU Emergency Service Department, Moi Teaching and Referral Hospital (MTRH), Uasin Gishu - County, between June 2016 and July 2016.
Methodology: Use of M/M/s queuing model to analyze ICU services using secondary data of MTRH emergency patients arrival and service rates together with estimated service cost of available 6 beds. Waiting cost estimated using formulated Modified Normal Loss function.
Results: With an average individual tolerance of $\tau = 0.083$ hrs and average response time of $\bar{x} = 0.083$hrs, the present scenario of 6 ICU beds in MTRH is operating at a service cost of Ksh 60 and patient queuing cost of Ksh 415.53 per hour. The length of the queue is 1.4 hr or approximately 34 patients per day. The optimum number of beds required for the facility to operate with zero response time and zero quality tolerance is 18 beds. This will facilitate the reduction of queuing cost to 153.98 and total cost to 333.98.
Conclusion: The current status of ICU emergency services at MTRH is costly to both the health facility and the patients. Individuals seeking such services may either opt to get similar services elsewhere at an opportunity cost of Ksh 641.39 per hour of delay. With 1.4 patients waiting in the queue every hour, this

**Corresponding author: E-mail: tisesko@yahoo.com;*

accumulates to 34 patients per day. Increasing ICU beds to 18 minimizes the length of the queue to 6 patients per day and queuing cost by 76% and reduces the total cost by 65%. This will reduce the financial burden of the patients and increase the chances of saving lives during emergency cases. These predictors, however, need further work and inclusion of related services to give a bigger and better picture of the facility.

Keywords: Queuing cost; service cost; normal loss function; individual tolerance; specification limit; optimum total cost.

2010 mathematics subject classification: 60K25, 90B22, 68M20, 91B06.

1 Introduction

A queue is a line formed by objects or people waiting to be served. In everyday life, queuing is inevitable. We find people queuing in the bank, hotel, bus stop, car wash, supermarkets, school, telephone boots, polling stations, traffic, airports, cafeterias, loading and offloading, hospitals just to mention but a few. Since queues are inevitable, the most important thing is to strike a balance between the length of the queue and the number of servers. An experiment on the fluctuating demand in telephone traffic was done by a Danish engineer named Erlang and a report addressing the delays in automatic dialing equipment was published, detailing equilibrium between the number of servers and the length of waiting time [1]. At the end of World War II, Erlang's early work was extended to more general problems and to business applications of waiting lines. This gave rise to the study of waiting lines, called queuing theory which is still used to date in customer service delivery [2].

In queuing theory, the three basic components of a queuing process are arrivals patterns, the actual waiting line and service facilities [2]. Customers arrive to the facility from an infinite calling population, with a random arrival pattern following Poisson process. Once customers arrive, they are served immediately if the server(s) is empty, or otherwise the customers wait in the queue for the next empty server. Mostly, the service is on a first come first serve (FCFS) basis although other methods like Service at Random Order (SARO) can be used. Preference service depending on the level of risk, urgency or the social, economic or political standing of the customer, and Hold on Line (HL) discipline, where important arriving customer takes the lead of the queue is rampant in many facilities. Customers who may feel to have waited for long in the queue can renege or balk and seek alternative equivalent services elsewhere, however, the queue length and waiting time depends on the traffic intensity, which is the ratio of arrival and service rates. The service discipline follows an exponential pattern, with individual service time variation due to different nature of the problems to be handled.

1.1 Queuing theory and health facility

A health facility provides a very essential service and the use of queues is crucial. The type of customers in a hospital is sick people whose lives are threatened and therefore needs urgent attention. Some of the customers queuing for services have life threatening conditions which require urgent attention and any delay in the queue means losing valuable time of saving life. Intensive Care Units (ICUs) are a critical component within hospitals, where patients are admitted when their vital functioning is compromised and their lives are in danger. With the introduction of motor bikes as a cheap and flexible means of transport, the number of fatal accidents increased and due to the availability of ambulance services, many accident patients can reach the hospital within a short time, thus increasing the arrival rate of customers who need urgent attention. Because some patients will remain in the ICU for a long time, and due to the nature and the cost of the life supporting machines necessary in the ICU, many hospitals aim at keeping a few but highly utilized units to address emergency needs. On the other hand, the demand for the service is high thus an equilibrium point needs to be determined for an optimum service delivery economical to the health facility and satisfactory to the customers.

Emergency case patients may be admitted to an ICU for intensive care immediately they arrive, or for postoperative care after an offensive operation. Emergency patients who need direct admission to the ICU may be rejected due to lack of space. This may lead to loss of lifers and poor image of the facility, not to mention loss of income which would be earned from the patient. For patients who need to be admitted to an ICU for postoperative care, lack of space in the ICU bed means the operation is either postponed or canceled. This cancelation may pose a severe health risk or have a major emotional impact on the patient. For the hospital, cancelation of operation may lead to unutilized operating room, which is equally costly to equip and thus a loss of resource capacity.

In most cases, patients are prepared for an operation long enough through procedures like testing blood pressure, blood levels and starving for over 6 hours. This means once an operation has been canceled, it is not possible to start another operation immediately. The availability of ICU beds is thus a highly important factor which reflects the service quality of the hospital. A wide variety of queuing models can be used in operations management, to help solve problems involving queue length, satisfaction of customers, idle servers and optimum service and waiting costs involved. In this paper, a multi-channel within finite calling population and first come first served discipline is adopted.

This model assumes that arrivals follow a Poisson Probability distribution and that service times is exponentially distributed and it is usually denoted by M/M/s: ∞/FCFS [3]. The cost of service versus the waiting costs is analyzed and equilibrium determined. The opportunity cost of customers waiting in the queue is highly individualized, but in this paper, it is assumed that the tolerance for the quality of service is normally distributed, and therefore the waiting costs follow a normal distribution. This paper will therefore focus on this ICU service, the cost in line waiting and the cost of service to optimize service delivery in Moi Teaching and Referral Hospital (MTRH).

The study sufficiently provides information to medical managers for decision making on the use of available limited resources to improve service offered to patients. The optimum ICU bed capacity is determined to ease congestion and at the same time minimize service cost.

2 Literature Review

Queues or waiting lines or queuing theory, was first analyzed by A.K. Erlang a Danish Engineer in the context of telephone facilities [1]. The body of knowledge that developed from it after came to be known as ”Queuing Theory”. It is widely practiced or utilized in industrial setting and management. Balancing the cost of providing services with the costs of customer waiting is the decision problem involved here. Use of queuing theory in health care is now utilized worldwide. Research has shown that queuing theory can be useful in real-world health care situations. McClain [4] reviews research on models for evaluating the impact of bed assignment policies on utilization, waiting time, and the probability of turning away patients. The use of queuing theory in pharmacy applications with particular attention to improving customer satisfaction is reviewed in [5] and the history on the use of queuing theory in health care facilities is presented in brief by [6]. However, it provides no description of the applications or results. Green [7] presents the theory of queuing as applied in health care. 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. The researcher agrees with them that queuing theory is of valuable use in evaluating health care facilities and will use it to solve the problem at hand.

The use of queuing theory as an analytical tool to predict how particular health care configurations affect delay in patient service and health care resource utilization with the associated costs. Fomundam and Herrmann [8] 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 health care 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 [9] calculated the percentage of patients who leave an emergency

department without getting help using arrival rate, service rate, utilization, capacity. From this percentage, 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 resources (doctors, ICU beds, machines etc.) in order to reduce overall costs. The extension to include stochastic models was done by [10].

The arrival of patients into the health care facility will be random and will follow the Poisson distribution. According to Karlin, and McGregor [11], 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 [12] studied how an increase in patient arrival rate affected waiting times and queue length for an emergency radiology service. Many health care systems have a variable arrival rate though some models assume a constant arrival rate, but 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. Similar results in the banking sector were echoed by [13].

Siddhartan, Jones, and Johnson [3] proposed a priority discipline for different categories of patients and then a first-in-first-out (FIFO) discipline for each category. They found that the priority discipline reduces the average waiting time for all patients. However, while the wait time for higher priority patients reduced, lower priority patients endured a longer average waiting time. An emergency anesthetic department operating with priority queuing discipline was modeled by [14] with an interest in the probability that a patient would have to wait more than a certain amount of time to be served. Haussmann, [15] 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 [16] 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 [17], analyzed 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.

Due to the availability of many parallel servers, the M/M/s queuing model is deduced from the Karlin and McGregor [11] representation for the transition probabilities. This representation allows for the study of arrival of patients, the queue length, the waiting in line cost and service cost. These will then enable us to determine the equilibrium to optimize service and reduce costs. Kembe [18], analyzed 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 over utilization 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.

According to Keller and Laughhunn [19], the capacity of the health care facility can be good but there is need to redistribute in time to accommodate patient arrival patterns. Other optimization designs proposed an incremental analysis approach in which the cost of an additional bed is compared with the benefits it generates [20]. 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 [21] stated 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 health care modeling through queuing theory has developed and there now exist a multitude of studies.

A considerable body of research has shown that queuing theory can be useful in real-world health care situations, and some reviews of this work have appeared. McClain [4], reviewed research on models for evaluating the impact of bed assignment policies on utilization, waiting time, and the probability of turning away patients. Gorunescu, McClean and Millard [22] 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 promptly and their survival rate is increased. 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 [23]. This is a special case, where the queue length cannot be greater than zero, which is called a pure loss model. Green [7] 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. Whitt [24] surveyed and developed time-varying queuing networks that help in determining the number of physicians and nurses required in an emergency department. 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.

3 Materials and Methods

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 three months was collected form Moi Teaching and Referral Hospital (MTRH) and spreadsheet software used to analyze the data. In this study, the M/M/s model was used to analyze the utility and cost optimization in the ICU health facility of MTRH. The following are the model characteristics and assumptions.

3.1 Model characteristics and assumptions

MTRH is a level five hospital serving more than 10 counties. The neighbouring health facility of the same standards is 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. 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 should be less than average service rate. This is necessary to ensure that the queue would not eventually grow infinitely.
(vii) Servers here represent doctors, beds, theater, 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 channels. This model could apply to many qualitative analysis of different situations. Some of the physical examples that apply include, a telephone booth 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 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 request 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): (∞/FCFS) is chosen to model the dynamics of an emergency medical service with respect to utility of ICU resources.

3.2 Model flow chart

Following the characteristics of the hospital emergency service, and the assumptions of the model, the following flow chart represents the components of a hospital queuing system. The system is illustrated to include servers, one queue and a general ward facility for recuperating patients. In this study, the patients admitted directly to the general ward are not considered to be in the queue, and those transferred from ICU to general ward are assumed to have left the system even when still admitted in the ward. Also, a patient admitted in the general ward who may require ICU services is assumed that the patient will join the queue for the services.

3.3 Model description (M/M/s): (∞/FCFS)

In (M/M/s) queuing model, it is assumed that the arrivals of patients follow a Poisson probability distribution at an average arrival rate of λ per unit of time. It is also assumed that arriving customers are served on a first-come, first-served (FCFS) basis by any of the empty servers with service times distributed exponentially with an average rate of μ per server per customer. With s number of servers, the average length of service time is $\frac{1}{s\mu}$.

If there are n patients in the queuing system at any point in time, then the following two cases may arise; Case I: That $n < s$, hence there will be no queue and $(s - n)$ number of servers will be idle. Case II: If $s \geq n$ then all servers will be busy and the maximum number of customers in the queue will be $(n - s)$.

Let p_o be the probability that there are no customers in the system, p_n be the probability of having 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 spend in the system, then; if λdt is the probability that an arrival enters the system between time t and time $t + dt$ interval, then $1 - \lambda dt$ is probability that no arrival enters the system within interval or dt time units. Also let μdt be the probability of one service completion between t and $t + dt$ time interval. Using $p_{n+i}(t);\ i = 0, 1, 2, \ldots$ as the transient state probability of exactly $n + i$ customers in the system at time t, and assuming the system started its operation at time zero, then $p_{n+i}(t + dt);\ i = 0, 1, 2, \ldots$ is the transient state probability of exactly $n + i$ customers in the system at time $t + dt$. As a property of the Multi-channel model, it is 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 $n = s - 1$. By discrete method and starting from when $n = 0$, the number of clients in the next unit of time is equal to the accumulation rate multiplied by the initial population, defined as;

$$p_1 = \frac{\lambda}{\mu} p_0 \tag{1}$$

When n lies between 1 and $s\ -\ 1$, all customers arriving will be immediately served and n channels out of s will be busy. The value of $p_n(t\ +\ dt)$ can occur in three exclusive and exhaustive ways and by considering the steady state of the system, these are obtained to be;

$$\lambda p_{n-1} -\ (\lambda + n\mu)p_n + (n\ +1)\mu p_{n+1} = 0;\ \ 1 \leq\ n \leq\ s\ -\ 1 \tag{2}$$

Putting $n\ = 1$ in equation (2), and using Equation (3.1), we get

$$p_2 = \frac{1}{2!}\left(\frac{\lambda}{\mu}\right)^2 p_0$$

and the recurrence relation for any value of $1 \leq n \leq s\ -\ 1$ is given in general by,

$$p_n = \tfrac{1}{n!}\left(\tfrac{\lambda}{\mu}\right)^n p_0 \tag{3}$$

In case $n \geq s$, we begin with when $n = s - 1$, substituting it in equation (2), we get

$$p_s = \tfrac{1}{s\mu}[\lambda + (s-1)\mu]p_{s-1} - \left(\tfrac{\lambda}{s\mu}\right)p_{s-2} \tag{4}$$

Now from Equation (3), $p_{s-1} = \tfrac{1}{(s-1)!}\left(\tfrac{\lambda}{\mu}\right)^{s-1} p_0$ and $p_{s-2} = \tfrac{1}{(s-2)!}\left(\tfrac{\lambda}{\mu}\right)^{s-2} p_0$ or in general, $p_s = \tfrac{1}{s!}\left(\tfrac{\lambda}{\mu}\right)^s p_0$. Similarly, when $n = s\ +1$, substituting in Equation (2) and simplifying, we obtain

$$p_{s+1} = \frac{1}{s.\,s!}\left(\frac{\lambda}{\mu}\right)^{s+1} p_0, \quad p_{s+2} = \frac{1}{s^2.\,s!}\left(\frac{\lambda}{\mu}\right)^{s+2} p_0, \ldots$$

which in general, for $n \geq s$,

$$p_n = \tfrac{1}{s^{n-s}.s!}\left(\tfrac{\lambda}{\mu}\right)^n \times p_0 \tag{5}$$

Thus equation (1), (4) and (5) give the value of p_n for $n\ = 0, 1 \leq\ n \leq\ s\ -\ 1$ and $n \geq s$. We now need to find the value of p_0 in terms of s, μ and λ. 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;

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

or

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

Replacing the first p_n using Equation (3) and replacing the second p_n using Equation (5), we obtain,

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

which with little algebra simplifies to

$$p_0 = \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \times \left(\frac{s\mu}{s\mu-\lambda}\right)\right]^{-1} \tag{6}$$

Now the other properties of the multi-channel system can be found using the equations in $\left((1)-(6)\right)$ as follows;

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

$$L_s = \frac{\lambda.\mu\left(\frac{\lambda}{\mu}\right)^s}{(s-1)!(s-2)^2}\, p_0 + \frac{\lambda}{\mu} \tag{7}$$

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

$$L_q = \left(\frac{\lambda.\mu\left(\frac{\lambda}{\mu}\right)^s}{(s-1)!(s\mu-\lambda)^2}\right) p_0 - \frac{\lambda}{\mu} \tag{8}$$

Using Equations (7) and (8), we determine the satisfaction of patients, using the parameter accounting for the average time a customer spends in the system defined as,

$$W_s = \frac{L_s}{\lambda} = \left(\frac{\mu\left(\frac{\lambda}{\mu}\right)^s}{(s-1)!(s\mu-\lambda)^2}\right) p_0 + \frac{1}{\mu} \tag{9}$$

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

$$W_q = \frac{L_q}{\lambda} = \left(\frac{\mu\left(\frac{\lambda}{\mu}\right)^s}{(s-1)!(s\mu-\lambda)^2}\right) p_0 - \frac{1}{\mu} \tag{10}$$

with the probability of having to wait given by the proportion defined in form of a probability as

$$p(n \geq s) = \left(\frac{\mu\left(\frac{\lambda}{\mu}\right)^s}{(s-1)!(s\mu-\lambda)^2}\right) p_0 \tag{11}$$

Under normal circumstances, no patient arrives to a health facility and finds no queue. This happens when the service rate μ is faster than the arrival rate λ. 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 $s -$(average number of customers served). The utilization rate of the servers is defined by $\rho = \frac{\lambda}{\mu}$ and thus the efficiency of M/M/s model is obtained from the traffic intensity,

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

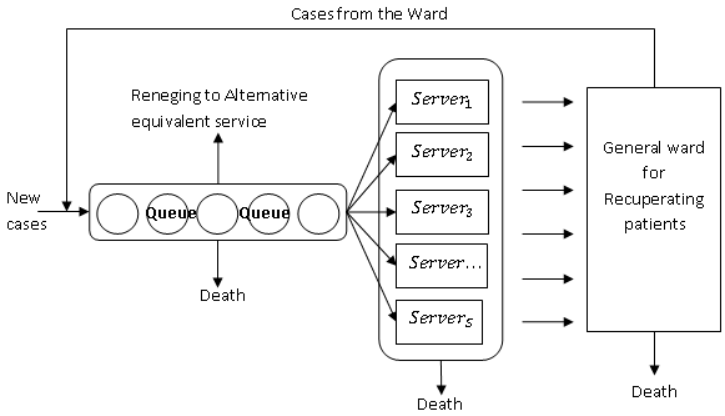

Fig. 1. Flow chart showing the major components of queuing system
Source: Author

4 Introducing Costs into the Model

In order to evaluate and determine the optimum number of servers required in the system, two opposing costs must be considered in making decisions: (i) Service costs (ii) Waiting costs. The first involves the cost incurred in the provision of desired service, here called service cost denoted by ($SC = C_s$). Service cost is directly incurred while providing services and it includes salaries paid to employees, cost of facilities, equipment and tools used, cost of service space, rent, supplies, ICU facilities, beds, electronics, doctor's fees, support staff allowances, oxygen, theater cost just to mention but a few. Waiting cost on the other hand entails the opportunity cost incurred by the customer due to waiting for service. It includes cost of losing life due to waiting, cost of getting equivalent service elsewhere, quality or value of time wasted and other costs associated with accepting incomplete and unsatisfactory services. This cost is referred to as service waiting cost and denoted by $WC = W_s$.

It is the wish of every customer to be given individual attention using the best equipment by qualified and experienced servant promptly. This is what they define as quality service. On the other hand, service provider on the other hand wishes to use minimum operation costs to provide services so as to maximize profit. This means customers will have to stand in long queues waiting to be served. Analysis of these costs helps in finding equilibrium point between the increased costs of providing better service and decreased waiting costs of customers.

In a hospital facility, the most crucial time is the waiting time in the queue before service starts denoted by W_q . Most patients belief that they are safe as soon as they see a doctor, even before treatment commence, but possibly first aid and pre-treatment tests conducted which include but not limited to checking blood reassure, body temperature, resuscitation, oxygen supply, rehydration, stopping bleeding or any other measure which reduces the risk that the patient is facing, and making them out of danger. Therefore, time in the queue would be the preferred characteristic to measure quality specification and thus used to estimate waiting and service cost. The waiting in line time for (M/M/s) model is defined in Equation (10) as $W_q = \frac{L_q}{\lambda}$, and using this relation, the expected cost of queuing in the system per person per unit time is the product,

$$C_q = \lambda W_q C_w = L_q C_w \tag{12}$$

Denote the expected service cost incurred by the health facility by C_s. Then the total service cost is given by

$$E(SC) = sC_s \tag{13}$$

where s is the number of servers. Using Equation (12) and (13), the total cost is obtained by adding the service cost and the waiting cost to yield,

$$TC = sC_s + C_q \tag{14}$$

4.1 Loss function for waiting lines

The measure of quality, as related to both the product and the service, is often dif cult to precisely quantify because of different perspectives of individuals, but quality involves short waiting time, cleanliness, satisfactory service, affordable and friendly. A low level of service may be inexpensive, in the short run but the service provider may incur high cost of customer dissatisfaction such as lose of future business, lose of a potential sale, development of poor reputation, loss of goodwill and increased competition by firms in the same industry. Deviation from the expected quality of service leads to a situation where the client incurs opportunity cost. The level of cost incurred can be determined by a loss function which links the cost and the level of deviation from the expected standards. The following two loss functions have been used previously in deter mining the opportunity cost a customer incurs when the product or service fails to meet the target specification value. These are the traditional loss function and the Taguchi loss functions discussed below.

4.1.1 Traditional loss functions

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 Fig. 2. 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 Fig. 3 was formulated using a quadratic function [25]. Principles of formulating Taguchi loss function assumes that, there is no cost incurred by the service providing organization or by the customer unless the product or service goes beyond its Upper or Lower Specification Limits, (USL or LSL).

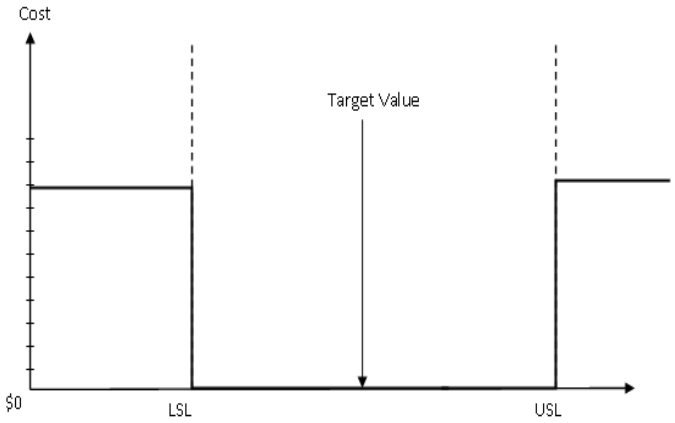

Fig. 2. Traditional loss function showing USL and LSL
Source: Author

4.1.2 Taguchi loss functions

The Taguchi Loss Function takes a different perspective on when the cost of poor quality are incurred. Taguchi theorized that rather than incur costs beginning at two finite points that are $\pm$ 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 [25]. 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 T.

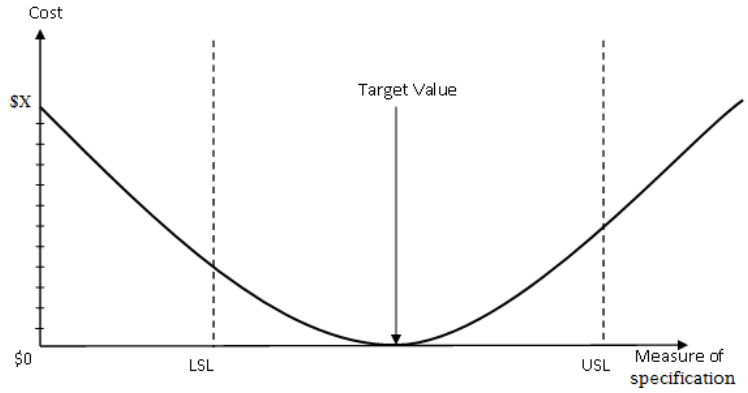

Fig. 3. Taguchi function showing calculation of loss function
Source: Author

In contrast with traditional models, the Taguchi Loss Function describes the failure to meet desired specification cost C incurred as;

$$C = \begin{cases} 0, & if\ LSL \geq m \geq USL \\ K(m-T)^2, & otherwise \end{cases} \tag{15}$$

where T is the target specification, m is the unit measure of quality specification and the constant K is determined from the cost of rejecting the item at a specification limit from the relation,

$$K = \frac{R}{(USL - T)^2},$$

where R is the cost or rejecting the item at specification limit. Clearly, the Taguchi loss function is a quadratic function which hits the zero cost line if the specification limit is equal to the target value $L = T$. It also has a uniform gradient for all customers and infinite cost on extreme deviation from the target value. This model does not put into account individual tolerance differences on levels of satisfaction of the item or service specification.

4.1.3 Modified normal loss function

In this study, it is assumed that individual preference is normally distributed with a mean of $\bar{x}$ and a standard deviation of σ. A normal distribution table is used to determine the proportion of the cost incurred if the measure of product or service quality deviates from the nominal specification value. In this case, the loss function $f(x)$ has a graph similar to an inverted normal curve, but with a gap of 2τ in between as shown in Fig. 4.

The cost function $f(x)$ is equivalent to a one sided normal distribution function which assigns a numerical value proportional to the amount at which the quality of service or product deviates from the individual specification target. The individual target quality will be a range of values in the interval $I := -\tau \leq T \leq \tau$. If the actual quality of the product or service $x = m$ is outside the interval I, the customers will start incurring costs due to dissatisfaction. The area under the curve $f(x)$, the x-axis and the lines $x = USL$ and $x = m$ measures the proportion of opportunity cost an individual will incur due to unsatisfactory standards.

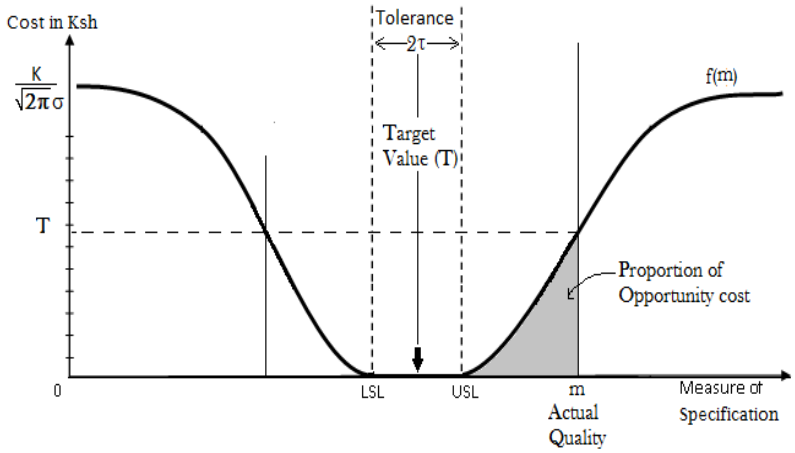

Fig. 4. Normal function showing calculation of loss function
Source: Author

The cost function is defined by,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}-\tau}{\sigma}\right)^2} \tag{16}$$

where $\bar{x}$ is the mean quality specification value, which is equivalent to the standards provided by the service provider and $\backslash sigma$ models the quality specification deviation, whileτ models the individual tolerance levels and x is the measure of the actual quality specification.

4.1.4 Tolerance and opportunity cost

The opportunity cost incurred due to delay of service or poor quality of products is inversely proportional to the individual level of tolerance. The lesser the tolerance, the higher the opportunity cost. Quality tolerance is hereby defined as the measure of the deviation from quality nominal specification without incurring any opportunity cost. This should be a natural stretch without any influence or duress. This characteristic is measured by the parameter τ, which accounts for individual differences on preferences or expected standards. The difference in individual preferences or tastes or tolerance is mainly due to individual lifestyle, social status, occupation, financial status, cost of alternative similar service elsewhere, urgency of the required service, risk associated with delay of the required service, purpose of the product required, just to mention but a few.

Let C_w be the cost of rejection at the specification limit $x = m$ and let T be the mean target specification value with individual level of tolerance of $\tau \geq T$. Then the actual cost of rejection incurred by the customer satisfies the condition

$$C_w = \begin{cases} Kf(m), & if \ \ m > |\bar{x} + \tau| \\ 0, & if \ \ m < |\bar{x} + \tau| \end{cases} \tag{17}$$

where K is the maximum opportunity cost incurred as $m \to \pm\infty$ and $\bar{x} + \tau = SL$ (Specification Limit).

4.1.5 Estimating waiting cost

The cost of waiting of an individual patient is estimated to be equal to the cost of the next best alternative opportunity, and it is proportional to length of time delay. The ICU services are very essential services which

save lives and the most important parameter of determining the quality of service is the response time $\bar{x}$ of the paramedics. Using time as the measure of quality specification, the target of every client is to be served at once without waiting. This means the value of $\bar{x}$ in our model is zero. Since we are using time t as a measure of quality, the parameter x in (12) will represent time. The new function will therefore be given by;

$$f(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{t-\tau}{\sigma}\right)^2} \tag{18}$$

The waiting cost therefore is determined using equation (14) as

$$C_q = Kf(t) \tag{19}$$

In order to use Standard Normal Tables, the observed values of service time x can be standardized using the relation $z = \frac{(t-\tau)}{\sigma}$ to obtain a standard normal distribution, which can be read directly from the tables.

Example

A server provides emergency services with an average target time of $\bar{x} = 0$ and a standard deviation of $\sigma = 15min$. Consider an individual who can incur a maximum opportunity cost of $K = Ksh500$ if service is delayed. If the individual has a tolerance of $\tau = 15min$ and the service was provided at $t = 40min$, then the waiting cost of the individual will be;

The standard score of $z = \frac{(40-0-15)}{15} = 1.667$

From the Normal Tables, the area under the curve $P(|z| = 1.67) = 0.9050$

The waiting cost will be $C_w = kf(t) = 500 \times 0.9050$

$$C_w = Ksh452.50$$

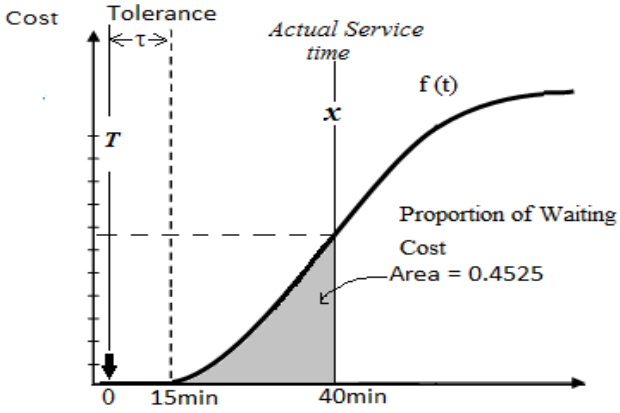

Fig. 5. Opportunity cost incurred
Source: Author

4.2 Queuing cost analysis

In this section, queuing costs is simulated in order to determine the optimum cost of the facility in relation to service cost and waiting costs. From section 4, using Equation (14) and Equation (19), we obtain the total cost as;

$$TC = sC_s + Kf(t) \quad (20)$$

4.3 Data analysis

The following data was obtained from MTRH showing the bed occupancy, or number of servers and service and arrival rates of the patients to the ICU. Using MATLAB, simulation was done to analyze the effects of varying the individual tolerance and response time. Computation of Queuing parameters and costs were done using MTRH ICU data presented in Table 1 below.

Table 1. Data showing values of parameters relating to MTRH ICU services

Item	Description	Symbol	Value
1	Number of ICU beds	s	Variable
2	Arrival rate of patients per unit time	λ	7
3	Service rate of each server	μ	2
4	Average expected response time (in hours)	$\bar{x}$	0.833
5	Standard deviation of service response time (in hours)	σ	0.083
6	Individual waiting time tolerance	τ	Variable
7	Maximum average waiting cost per individual (Ksh/hr)	M	800
8	Average service cost of each server (in Ksh) per hour	K	100
9	Measure of service quality specification	m or t (time)	Variable
10	Individual target service time	$\bar{x} + \tau$	Variable
11	Total number of customers arriving in the facility	n	$n > s$

Using the data in Table 1, the following costs are generated from MATLAB simulation. Simulation for 5 beds to 32 beds results is presented in Table 2. This is done with tolerance level of $\tau = 0.083$ hours and a response time of $\bar{x} = 0.083$ hours. With service time standard deviation of $\sigma = 0.0162$ hours, we obtain the optimum queuing cost of Ksh 129.52 per hour and the desired number of servers as $s = 14$ (See Table 2).

Table 2. Simulation results for non-zero tolerance $\tau = 0.083$, and $\bar{x} = 0.083$

s	f	L_q	L_s	W_s	W_q	C_q	C_s	TC
5	0	2.3333	5.8333	0.83333	0.33333	0	50	50
6	**0.37101**	**1.4**	**4.9**	**0.7**	**0.2**	**415.53**	**60**	**475.53**
7	0.42463	1	4.5	0.64286	0.14286	339.7	70	409.7
8	0.46583	0.77778	4.2778	0.61111	0.11111	289.85	80	369.85
9	0.49816	0.63636	4.1364	0.59091	0.090909	253.61	90	343.61
10	0.52405	0.53846	4.0385	0.57692	0.076923	225.74	100	325.74
11	0.54515	0.46667	3.9667	0.56667	0.066667	203.52	110	313.52
12	0.56265	0.41176	3.9118	0.55882	0.058824	185.34	120	305.34
13	0.57736	0.36842	3.8684	0.55263	0.052632	170.17	130	300.17
14	0.58988	0.33333	3.8333	0.54762	0.047619	157.3	140	297.3
15	**0.60066**	**0.30435**	**3.8043**	**0.54348**	**0.043478**	**146.25**	**150**	**296.25**
16	0.61003	0.28	3.78	0.54	0.04	136.65	160	296.65
17	0.61825	0.25926	3.7593	0.53704	0.037037	128.23	170	298.23
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
28	0.66854	0.14286	3.6429	0.52041	0.020408	76.405	280	356.4
29	0.67115	0.13725	3.6373	0.51961	0.019608	73.695	290	363.69
30	**0.67357**	**0.13208**	**3.6321**	**0.51887**	**0.018868**	**71.17**	**300**	**371.17**
31	0.67584	0.12727	3.6273	0.51818	0.018182	68.812	310	378.81
32	0.67795	0.12281	3.6228	0.51754	0.017544	66.606	320	386.61

In absence of tolerance, simulation results show an increase in the queuing cost. This is expected because absence of tolerance increases the length of waiting time in the queue which attracts costs. In this scenario, at $\tau = 0$, queuing cost increased to $C_q = 145.93$ and an optimum number of servers of $s = 18$. This will allow for having an extra idle server/bed and have traffic intensity $\rho \leq 1$ to handle emergencies. These results concur with the findings of [13]. The combined graph for the two scenarios is depicted in Fig. 6.

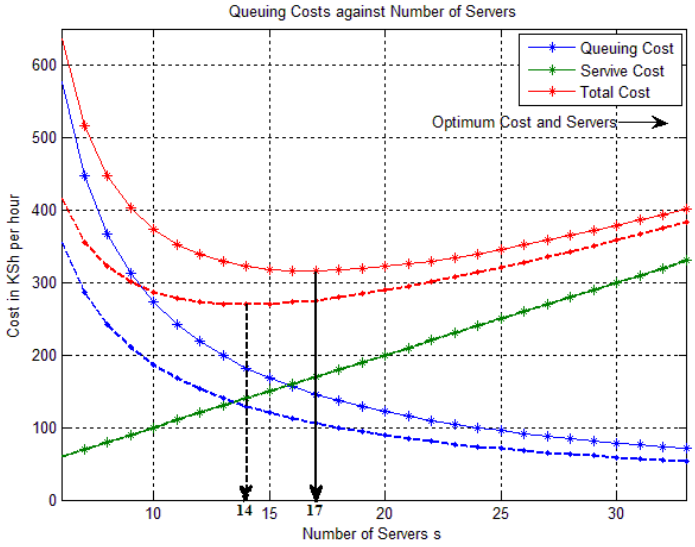

Fig. 6. Queuing costs against number of servers

The response time affects the cost of queuing if the level of tolerance is shorter than the response time. In any case, the length of delay which attracts cost is calculated at $W_q - \tau$. Using the normal curve with $N(\bar{x}, \sigma) = N\ (0.0083, 0.0083)$, the corresponding probability of waiting in the queue is $W_q = 0.7036$ which gives the proportion of total cost of waiting. The Queuing cost is therefore calculated using Equation (19) as, $C_q = Kf(W_q) = 0.7036 \times \lambda \times W_q \times 800 = 562.88$.

5 Conclusions

In the analysis done in section 4, the queuing characteristics at Moi Teaching and Referral Hospital (MTRH) in Eldoret, Uasin-Gishu County was analyzed using M/M/s queuing model. Waiting and Service costs were introduced into the model in order to determine the optimal service level. With the support from the government, the target was to determine the optimum number of servers which will reduce the patient queuing cost thus satisfying the consumers and save lives. Incorporating tolerance in the loss function, the proportion of customer's willingness to meet the waiting costs due to delayed service was used to determine the queuing cost. From the current scenario of 6 ICU beds, the operation cost in absence of delay (tolerance) is estimated at, $C_s = 60$ and $C_q = 415.53$, with $W_q = 1.4$ people per hour. Since ICU operates 24 hours a day, this translates to a service and queuing cost of $Ksh.\,1{,}440$ and $Ksh.\,9{,}972.72$ per day. If the first set of patients admitted to ICU spends an average of 4 days in bed, the total accumulated people waiting in the queue will be 134.4 patients with a total opportunity cost of $Ksh.\,55{,}847.23$. The following are the results of the model analysis.

a) In subsection (3.3), M/M/s queuing model is analyzed to determine the number of customers in the system L_s , length of the queue L_q , waiting time in the system W_s and waiting time in the queue W_q.

b) In subsection (4), the total cost $TC = sC_s + C_q$ was represented in terms of the service cost C_s and the queuing cost C_q.
c) In the sub subsection (4.1.3), a modified normal loss function was formulated. This probability function $f(x)$ use the individual level of tolerance τ to determine the proportion of loss the individual incur due to waiting in the queue. This loss function modifies the waiting cost to $C_q = Kf(t)$.
d) In the subsection (4.3), data from MTRH is simulated using equations ((3.7)-(3.10), (4.5)-(4.7)) to determine the optimum total cost for different number of servers (beds). These simulation results are shown in Table 2.
e) From simulation results in (d) above, it is shown that in absence of tolerance $\tau = 0$, the facility operating currently at $s = 6$ beds has ≈ 34 patients waiting in the queue in a day each incurring a waiting cost of $Ksh\ 641.39$ per hour. If the beds are increased to an optimum number $s = 18$, the queuing cost will reduce to $Ksh\ 153.98$, which is a 76% reduction. If the patients allow a tolerance of $\tau = 0.083$ hrs, the optimum beds will reduce to $s = 15$ with a corresponding queuing cost of $Ksh\ 146.25$ representing a 65% reduction from the current scenario. These optimum points are illustrated in Fig. 6.

Following the description of results above, data analysis indicates that there is a strong need to acquire more ICU beds to a minimum of 18. This is less than 30 units proposed by the current CEO in an article which appeared in the Kenya Daily Nation dated Wednesday 27th July 2016. If the CEO's proposal is implemented, the facility will operate at $s = 30,\ C_q = 87.91,\ C_s = 300,\ TC = 387.91$. This will reduce the customers queuing cost by 86% and reduce the total operation service cost by 45%.

This analysis can however be extended to include related facilities and departments like the theater, wards, endoscopy, dialysis, RMI, X-ray, CT scan services and number of doctors, just to mention but a few. This will give an overall performance of the entire facility. The data used was collected for three weeks between June and July 2016. Reliability of analytic results will improve if data is collected for a longer period of time.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Erlang AK. The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik B. 1909;20(33-39):16.

[2] Bhat UN. An introduction to queueing theory: Modeling and analysis in applications. Birkhäuser; 2015.

[3] Siddharthan K, Jones WJ, Johnson JA. A priority queuing model to reduce waiting times in emergency care. International Journal of Health Care Quality Assurance. 1996;9(5):10-16.

[4] McClain JO. Bed planning using queuing theory models of hospital occupancy: A sensitivity analysis. Inquiry. 1976;13(2):167-176.

[5] Nosek RA, Wilson JP. Queuing theory and customer satisfaction: A review of terminology, trends, and applications to pharmacy practice. Hospital Pharmacy. 2001;36(3):275-279.

[6] Preater J. Queues in health. Health Care Management Science. 2002;5(4):283-283.

[7] Green L. Queueing analysis in healthcare, in patient flow: Reducing delay in healthcare delivery. Springer. 2006;281-307.

[8] Fomundam S, Herrmann JW. A survey of queuing theory applications in healthcare; 2007.

[9] Broyles JR. Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression. In IIE Annual Conference. Proceedings. Institute of Industrial Engineers-Publisher; 2007.

[10] Armony M, et al. On patient flow in hospitals: A data-based queueing-science perspective. Stochastic Systems. 2015;5(1):146-194.

[11] Karlin S, McGregor J. Many server queueing processes with poisson input and exponential service times. Pacific J. Math. 1958;8(1):87-118.

[12] Rosenquist CJ. Queueing analysis: A useful planning and management technique for radiology. Journal of Medical Systems. 1987;11(6):413-419.

[13] Bakari H, Chamalwa H, Baba A. Queuing process and its application to customer service delivery (A case study of Fidelity Bank Plc, Maiduguri). International Journal of Mathematics and Statistics Invention (IJMSI). 2014;2(1):14-21.

[14] Taylor TH, Jennings AMC, Nightingale DA, Barber B, Leivers D, Styles M, Magner J. A study of anaesthetic emergency work. Paper 1: The method of study and introduction of queuing theory. British Journal of Anaesthesia. 1969;41:70-75.

[15] Haussmann RD. Waiting time as an index of quality of nursing care. Health Services Research. 1970; 5(2):92.

[16] McQuarrie D. Hospitalization utilization levels. The application of queuing. Theory to a controversial medical economic problem. Minnesota Medicine. 1983;66(11):679.

[17] Worthington D. Queueing models for hospital waiting lists. Journal of the Operational Research Society. 1987;38(5):413-422.

[18] Kembe M, Onah E, Iorkegh S. A study of waiting and service costs of a multi-server queuing model in a specialist hospital. International Journal of Scientific & Technology Research. 2012;1(8):19-23.

[19] Keller T, Laughhunn D. An application of queuing theory to a congestion problem in an outpatient clinic. Decision Sciences. 1973;4(3):379-394.

[20] Young J. Estimating bed requirements. A queueing theory approach to the control of hospital inpatient census. John Hopkins University, Baltimore. 1962;98-108.

[21] Green L, Kolesar P. The pointwise stationary approximation for queues with nonstationary arrivals. Management Science. 1991;37(1):84-97.

[22] Gorunescu F, McClean SI, Millard PH. A queueing model for bed-occupancy management and planning of hospitals. Journal of the Operational Research Society. 2002;53(1):19-24.

[23] McManus ML, et al. Queuing theory accurately models the need for critical care resources. The Journal of the American Society of Anesthesiologists. 2004;100(5):1271-1276.

[24] Whitt W. What you should know about queueing models to set staffing requirements in service systems. Naval Research Logistics (NRL). 2007;54(5):476-484.

[25] Fink R, Gillett J. Queuing theory and the taguchi loss function: The cost of customer dissatisfaction in waiting lines. International Journal. 2006;17.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/16552