Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**
Abbreviated Key Title: Sch. J. Phys. Math. Stat. **ISSN 2393-8064 (Online)**

(An International Publisher for Academic and Scientific Resources)

Queueing Models in Healthcare with applications to a General Hospital in Zimbabwe

Romeo Mawonike[1*], Thompson Mahachi[2]
[1,2]Department of Mathematics and Computer Science Great Zimbabwe University, P O Box 1235, Masvingo, Zimbabwe

***Corresponding author**
Romeo Mawonike

Article History
Received: 25.03.2018
Accepted: 03.04.2018
Published: 30.04.2018

DOI:
10.21276/sjpms.2018.5.2.12

Abstract: In healthcare sector, quality services come with a compromise of devoting more resources e.g. labour force, waiting space, efficient laboratory equipment, etc. Few workforce result in prolonged and sluggish queues which are life threatening especially to accident ill patients. Waiting on a queue is not usually interesting, but reduction in this waiting time usually requires planning and extra investments. Still, emergency departments and intensive care units are among the most intricate and expensive of all medicinal resources, and hospital authorities are mandated to meet the demand for intensive care services with suitable capability. This study seeks to address the congestion of patients flow in acute departments (Emergence department and Intensive care unit ward) from our local hospitals by applying analytical queueing models to the situation. However, our models in this paper only address the waiting queues and waiting space as main challenges affecting patients in Emergence department (ED) and Intensive Care Unit (ICU) department. Results show that more doctors are required in both ED and ICU to serve patients to reduce queues and serve more lives. In addition, more waiting spaces should be created to accommodate more patients to avoid "blocking" to patients.
Keywords: Waiting lines; patients; doctors; waiting space; general hospital.

INTRODUCTION

Health care systems and hospitals in particular are our life jackets which hold back our lives every day. The rich and the poor need these facilities in time of trouble and even when things seem okay. Therefore, adequate resources from both government and private institutions must be allocated to this sector to enhance timely deliverance of the crucial services to reduce deaths and prolonged queues by patients. To achieve this, the management should be well equipped with the knowledge of queueing models and queueing systems of the institution(s). A queueing model is a mathematical description of a queueing system which makes some specific assumptions about the probabilistic nature of the arrival and service processes, the number and type of servers and the queue [1].

Queues have existed long back historical records could not account for but mathematical analysis of queues emerged as late as in 1900 century [2]. Queueing theory was developed by A. K. Erlang in 1904 to assist in determining the capacity requirements of the Danish telephone system [3]. Waiting in a queue is not usually interesting, but reduction in this waiting time usually requires planning and extra investments. When it comes to healthcare, queueing models can be helpful in allocating the appropriate number of beds, the level of staff and medical equipment as well as making informed decisions about how to allocate resources to different departments [4].

We rely on medical centres which provide defensive care and treat our illnesses, injuries and diseases. Truthfully, health care is possibly the arena determinant of people's quality of life and prolonged existence [5]. Healthcare systems have been challenged in recent years to deliver high quality services with limited resources and these resources are becoming increasingly limited and expensive. Therefore, much care is needed on how to utilize them with regard to the number of services required by patients, not forgetting to emphasis on quality. Nevertheless, emergency departments (ED) and intensive care units (ICU) are among the most important and expensive of all medical departments. Hospital authorities are challenged to satisfy the demand for emergency and intensive care services with an appropriate capacity [6]. Recent research shows that in the ED and ICU patients experience longer waiting times to be admitted or diverted from a unit (as a bottleneck unit) as it reaches full capacity resulting in limited access of healthcare to the public, on the other hand increasing operational costs due to inefficiencies. If too much service is provided, more excessive costs are incurred while providing inadequate service capacity causes the waiting line to become terribly long.

The performance of a hospital can be measured in a variety of ways [7], that is; financially, clinical, quality and timely service delivery, operational, psychological and other societal dimensions. The definitive goal is to achieve an economic balance between the cost of service and the cost connected with the waiting for that service [8]. The number of refused admission at the ED and ICU is high and many patients are diverted or referred to mostly private hospitals.

Background

A General Hospital under study is a state owned institution with a mandate of delivering health services to a District community (Mashonaland East) in Zimbabwe. The General Hospital is located about 800 m from Harare – Beit bridge Highway in the Eastern side of Chivhu town, and has a projected catchment population of about 125 028, that is, according to Zimbabwe National Statistics Agency (ZIMSTATS) in its 2012- 2022 population projections. There is another nearby general hospital which is about 67 km away. Other clinics in the vicinity are 3.5 km, 28 km, 50 km and 32 km away respectively. However, most of the referrals are done at Harare General Hospital, Chitungwiza General Hospital and Parirenyatwa Referral Hospital due to lack of resources in small medical centers. The institution has a busy Acute Care unit and this is attributable to unceasing accidents occurring along Harare – Beitbridge highway which connects South Africa to Zimbabwe and other southern regional countries.

The General Hospital's everyday businesses and activities are quite numerous and complex to analyze, but for the purpose of this study, we only considered bed requirements for patients arriving through the Acute Care Unit. This study intends to analyze more fundamentally some aspects of the queueing network that has been developed to model patient flows in the General Hospital's Acute Care Department. Acute care Unit in this study is made up of two main Departments, that is, the Emergency Department and the Intensive Care unit, where severe medical conditions are attended for only a short period of time and at a crisis level. Many hospitals have acute care facilities with the goal of discharging the patient as soon as the patient is deemed healthy and stable. The rising population in this District has created a need to understand how hospital resources relate to the quality of service in acute care facilities. In this research, hospital resources are measured in terms of beds (waiting space), accompanied by the essential medical equipment and staff (physicians). To ensure an adequate level of access to care, it is important to examine future bed requirements and the number of physicians required to cater for the growing population. The accurate prediction of this count requires both the knowledge of future population demographics, which affects the demand for acute care services, and also an understanding of how the number of available beds affects access to care.

The general hospital in question has exhibited failures and challenges in queue management and to that effect the study intends to explore on the efficient queueing models that will improve the situation. After having noted the above mentioned concern, it was deemed worthwhile to understudy the queueing processes involved in the Emergency Department and the Intensive Care Unit among multiple compartments/departments that may be found at the hospital. Generally, queueing models have not been widely and usefully applied in most general hospitals in Zimbabwe. Ideas are there but there are still in infantry stages. Queueing analysis in these hospitals is based on Exploratory Data Analysis (EDA) rather than analytical queueing models.

The Basic Queueing System

A basic queueing system is a service system where *customers* arrive to a pool of parallel or serial *servers* and require some service from one or some of them. A server is a person or anything that provides the service. If all servers are busy upon customer's arrival, the customer joins a *queue*. The law that determines the order in which queued customers are served is called the queue *discipline*. The most commonly followed discipline is *first-in, first out* (FIFO) rule, but other disciplines are often used to enhance efficiency or reduce the delay for more priority customers. *Triage* is also used by hospital emergency rooms depending on the criticality of the patient's injury as patients arrive. That is, a patient with a broken rib or neck receives top priority over another patient with a small cut on the hand. This may be pre-emptive or non-pre-emptive, depending on whether a service in progress can be interrupted when a customer with a higher priority arrives. Usually, in most queueing models, we assume that the calling population has *infinitely* many customers who would require the service from time to time. A short summary of how the system evolves over time is demonstrated in figure 1.

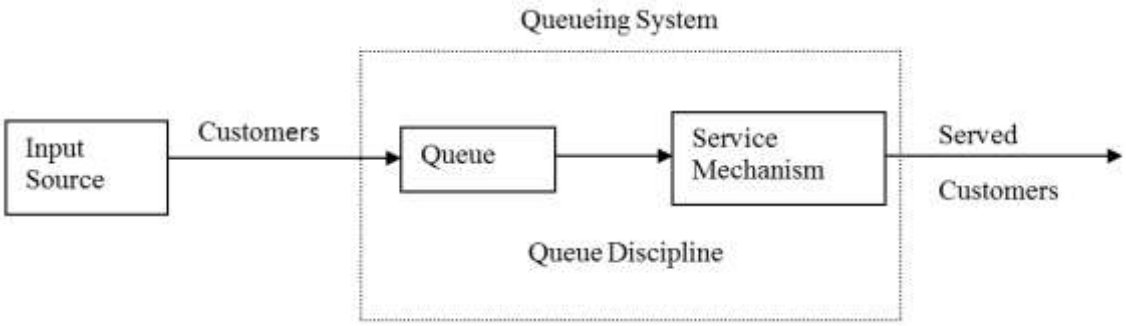

Fig-1: Structure of a general queueing system

The Poisson Arrival Assumption

In most queueing situations, we make assumption on the distribution of arrivals and that assumption says arrivals follow a *Poisson process*. This comes from the fact that the number of arrivals in any given time period has a Poisson distribution. This Poisson distribution is derived from arrivals that generate a counting process$N(t)$. We can show analytically that if customers arrive at random (independently from one another); the arrival process is a Poisson process. Because of this reason, the Poisson process is regarded the most random arrival process. For its prevalence and assumption of independent arrivals, the Poisson process is the most commonly used arrival process in modelling service systems. Poisson has been a good representation of unscheduled arrivals in many departments of the hospital including EDs, obstetrics and ICUs [9].

The Exponential Service Time Assumption

The most important property of the Exponential is that of being memory-less. It means that the time to the next arrival is independent of the time of the last arrival occurred. The property also leads to the fact that if the arrival process is Poisson, the number of arrivals in any given time interval is independent of the number in any other non-overlapping time interval. That is, the inter-arrivals follow an Exponential distribution. This also follows that service times are indeed Markovian (follow Exponential distribution).

Queueing Behavior

The behavior of patients while in queue and service encounters is very unpredictable. Some patients have a propensity to renege while in queue and this interrupts the queueing system. *Reneging* is the process of a patient entering the queue but deciding to leave the queue and the hospital unattended. In situations where the waiting line is long and inconvenient, patients are likely to balk. *Balking* is the process of a patient evaluating the queue and server system and registers displeasure of joining the waiting line. In both situations, the patient leaves the hospital, and may not return, and this is common in government hospitals where long queues are experienced. Therefore, those patients with reputable medical aid facilities would seek the attention of private doctors. In case of a multiple server system (more than one physician), patients may "jockey" for place between servers. Jockeying is a process where patients switch in between queues because of lack of satisfaction of oneself. In our study, we assume that a patient has patience, hence there is neither renege, balk nor jockey since these situations are difficult to model without simulation.

Definition of Terms

Emergency Department (ED): sometimes termed Emergency Room, this unit provides initial treatment to patients with a broad spectrum of illnesses and injuries, some of which may be life-threatening and requiring immediate attention.

Intensive Care Unit (ICU): is a specialized department used for intensive care medicine. Most patients arriving to the ICU are admitted from the ED. After their treatment, ICU patients are usually transferred to the medical unit for further care before discharge. This transfer, however, is possible only if a bed is available in the medical unit.

The service discipline: refers to the procedure for selection of units from the queue for service. The most commonly used disciplines in healthcare are FCFS and priority (pre-emptive and non-pre-emptive).

Related Literature

Patient inflow and outflow in hospitals has been studied lengthily. Interested researchers and readers are referred to papers [7-12] and the references therein. Most researchers tried to solve problems of queueing in ED ignoring the Internal Wards (IW) because of the complexity or non-accessibility of the data [7]. However, there are a few researchers who looked on broader issues e. g. [13] identify the main cause of ED congestion and variability. In the same manner, De

Bruin *et al*. [14] observe that refused admissions at the First Cardiac Aid are principally caused by unavailability of beds downstream the care sequence. These blocked admissions can be controlled via proper bed allocation along the care chain of Cardiac inpatients to sustain such allocations. Therefore, they proposed a queueing network model with parameters that were estimated from hospital data. Expanding the view Hall *et al*. [15], develops data based descriptions of hospital flows, starting at the highest unit level down to the specific sub-wards. Some researchers in this field of healthcare used queueing methods as optimization techniques in trying to identify the required number of beds needed in ED and ICU. De Bruin *et al*. [16] and Green [11], develop queueing models such as Erlang-C and loss systems to recommend bed distribution strategies for hospital wards. Time-varying queueing networks were developed in Green *et al*. [17] to assist in determining the number of physicians and nurses required in the ED. Some researchers do not apply analytical models of queueing theory to analyze the queueing system of a hospital, e. g. Armony *et al.* [7] apply Exploratory Data Analysis (EDA) in a large hospital setup taking into consideration the ED and the IW branches.

It is noted that M/M/c) queueing model can accurately model the flow of in-patient in hospital by determining the optimal number of beds needed in both the Intensive Care Ward (ICW) and Medical and Surgical Ward (MSW) so that the admission of patients into Emergency and Accident Department (EAD) is not affected [1]. In the same manner M/G/C queue was used with state dependent arrival rate to address the long- wait list problem, various management actions were applied such as increasing the number of beds or decreasing the mean service time through appropriate means [18,19]. Optimal bed allocation can be determined in the emergence cardiac patient hospital in-order to keep the fraction of refused admission [20-22]. Queueing models are of paramount in healthcare, they can also be applied to the congested patient flows in mental health systems [23], analyse hospital bed planning under peak loading [24], model the admission-and-discharge data of a specific ICU [25, 26] and determine the optimum number of nurses needed in an antenatal clinic to reduce the time spent by pregnant women in the queue and the system [27].

MATERIALS AND METHODS

Research Data

Data used in this research spans from January 2013 to December 2017. Two variables were considered that is; the arrival time and service time of patients at a general hospital. We have recorded patients' arrival times into ED, their departure times and mode of arrival (that is either independently or by ambulance). We have also recorded patients' arrival times into the ICU (either by being transferred from the ED, referred from other nearby hospitals/clinics or are from their homes) and their departure times. We only considered two departments (ED and ICU) because of time management, we hope next time we will consider all departments for the sake of completeness as well as doing much justice to the authorities for informed decision making. In ED we have two types of patients; the boarding and in-process patients. *Boarding patients* are those patients in ED awaiting hospitalization and *in-process patients* are those patients who have just arrived and receiving treatment or under evaluation [7].

We have noted that there is a small percentage (10%) of arriving patients being discharged after body examinations and tests. 90% of these patients are admitted into the hospital either straight into the ICU or to the ED for clinical treatment and check-ups. We give apologies that our models could not capture the 10% of these patients and operational costs are difficult to quantify and enumerate especially the waiting and service costs. The ED and ICU cannot handle more than 10 patients and 20 patients per unit time respectively.

Birth and Death Processes

In many real life applications, the state of the system sometime increases by one and other times decreases by one and no other transitions are possible [28]. Such a Markov chain $\{X_n\}$ is called a *birth and Death process*. Suppose we consider a counting process$N(t)$, for $0 \leq t < \infty$. $N(t)$ can be the number of patients waiting or in service at time *t*. We say that a system is in state E_j at time *t* if $N(t) = j$. The process $N(t)$ is a *birth and death process* if it obeys the following postulates. If at any given time *t* the system is in stateE_j, the conditional probability that during$(t, t + \delta t)$, the transition $E_j \rightarrow E_{j+1}, (j = 0, 1, 2, \dots)$ equals $\lambda_j \delta t + 0(\delta t)$ as $\delta t \rightarrow 0$, and the conditional probability of the transition $E_j \rightarrow E_{j-1}, (j = 1, 2, \dots)$ equals $\mu_j \delta t + 0(\delta t)$ as $\delta t \rightarrow 0$. The probability that during$(t, t + \delta t)$, the index j changes by more than one unit is $0(\delta t)$ as $\delta t \rightarrow 0$, (See Cooper [29]) for more details).

Applying the law of total probability, we write;

$$P\{N(t + \delta t) = j\} = \sum_{i=0}^{\infty} P\{N(t + \delta t) = j | N(t) = i\} P\{N(t) = i\} \quad (1)$$

Now it follows from the postulates that as $\delta t \rightarrow 0$,

$$P\{N(t + \delta t) = j | N(t) = i\} = \begin{cases} \lambda_{j-1}\delta t + 0(\delta t), i = j - 1 \\ \mu_{j+1}\delta t + 0(\delta t), i = j + 1 \\ 0(\delta t), |i - j| \geq 2. \end{cases} \quad (2)$$

Since we require from (1) and (2) and other postulates that,

$$P_j'(t) = \lambda_{j-1}P_{j-1}(t) + \mu_{j+1}P_{j+1}(t) - (\lambda_j + \mu_j)P_j(t), j = 0, 1, 2, \quad (3)$$

The coefficients $\{\lambda_j\}$ and $\{\mu_j\}$ are called birth and death rates respectively. When $\mu_j = 0$ for all *j*, the process is called a *pure birth process* (only arriving patients are allowed and no one is being treated) and when $\lambda_j = 0$ for all *j,* the process is called *a pure death process (*arriving patients are not admitted into and only admitted patients are treated and discharged) Solving the difference differential equation in (3) by induction, considering a special case of pure birth process with constant birth rate $\lambda_j = \lambda$ assuming that the system initially was in state E_0, then it is easy to verify that the counting process $N(t)$ is a Poisson process with mean λt. If we combine two processes that is birth process and death process, we come up with an ideal queueing system where patients arrive to a system and get service. We discuss some of the birth and death models below.

Models with Infinite Waiting Room

In this section we discuss two queueing models, the first one considers only one server (physician) and the second one considers two physicians serving one queue. In both models, a single queue is tolerated and no restricted number of arrivals is imposed.

M/M/1 Queueing Model

This is the simplest model in queueing theory, only one physician is available to attend to waiting patients. The waiting room is assumed to accommodate an infinite number of patients. There is no loss of generality to assume that arrivals follow a Poisson distribution with mean rate λ and service times follow an Exponential distribution with mean rate μ. The first can be proved using the distribution of counting processes where arrivals are independent events occurring within small intervals. The later can be proved using inter-arrival distribution of events. In healthcare facilities, it is unusual to find waiting rooms which can accommodate infinity number of patients especially in ED and ICU departments where waiting spaces are regarded as beds. Of course in medical wards, we can find infinity space since patients can wait on benches and chairs in a large room. We state without proof the steady-state probability distributions and measure of performance of this model in equations (4), (5, (6). The calculations can be derived from Chapman-Kolmogorov equations or an analysis of the embedded Markov chain at customer arrival and or departure epochs [2]:

$$p_d = 1 - e^{-\mu(1-\rho)t} \quad (4)$$

The expected average delay or average length of inpatient stay (ALOS) is given by $E(D) = \frac{1}{\mu(1-\rho)}$. The probabilities that a patient can wait for more than *t* time in the queue and or in the system are given by:

$$P(W > t) = e^{-\mu(1-\rho)t} \quad (5)$$

$$P(W_q > t) = \rho e^{-\mu(1-\rho)t} \quad (6)$$

M/M/c Queueing Model

This model, we consider more than one physician to attend to critically injured patients. Again, the waiting room accommodates infinite number of patients. Due to the increase in the number of accidents along the Beitbridge highway road, there is need to consider more than one doctor per each critical department. Assuming a Poisson arrival distribution with parameter λ and Exponential service distribution with parameter $c\mu$, the steady-state equations of the model are presented in (7) and (8).

$$p_0 = \left\{\sum_{n=0}^{c-1}\frac{\rho^n}{n!} + \frac{\rho^n}{c!(1-\frac{\rho}{c})}\right\}^{-1} \quad (7)$$

$$p_n = \begin{cases}\frac{\rho^n}{n!}p_0, 0 \le n \le c \\ \frac{\rho^n}{c^{n-c}m!}p_0, n \ge c\end{cases} \quad (8)$$

We also present the probability of waiting in (9) and (10).

$$P(W > t) = e^{-\mu t}\left\{1 + \frac{p_0\rho^c}{c!(1-\rho)}\left[\frac{1-e^{-\mu t(c-1-\rho)}}{c-1-\rho}\right]\right\} \quad (9)$$

$$P(W_q > t) = (1 - \sum_{n=0}^{c-1}p_n)(e^{-c\mu(1-\rho)t}) = p_d(e^{-c\mu(1-\rho)t}) \quad (10)$$

Models with Limited Waiting Room

Overcrowding in urgent clinics and hospitals is common in healthcare [2]. When the waiting room is small or can accommodate a limited number of patients, the queueing system becomes affected. Hence suitable models which can cater for limited number of patients in the system are employed. When the system is full to its capacity, only two options are modelled [2]. Firstly, Hospital authorities may limit the number of patients into the system including the one under service and new arrival(s) would not be allowed to enter until one leaves the system. Secondly, the system may be allowed to run as an infinite capacity, patients not accommodated in the system may decide to leave the system unattended and are referred to another clinic or hospital depending on the criticality of their situation, and hence customers are lost. We discuss two separate models with finite capacity assuming a Poisson arrival rate and Exponential

service rate. We prefer a Poisson arrival rate because it is flexible in both situations mentioned above.

M/M/1/K Queueing Model

This model has one physician to attend to a restricted number of patients because of limited space to accommodate the waiting patients. Some of the patients are lost after there have discovered a full capacity system. Stability is guaranteed here because the size of the queue cannot exceed the available spaces in the waiting room. In our case, this model is ideal as we have a limited number of beds in ED and ICU. The waiting time is reduced as the number of beds decrease but there will be again a decrease of revenue since most of the critical patients are referred to other hospitals. On the other hand, the increase in the number of beds against one doctor results in a prolonged service. This model takes into consideration of the lost customers by introducing $\lambda_{effective}$ in place of λ. We again state without proof the steady-state equations (11) and (12) for this model.

$$p_0 = \frac{1-\rho}{1-\rho^{K+1}}, \rho \neq 1 \tag{11}$$

$$p_n = \left(\frac{1-\rho}{1-\rho^{K+1}}\right)\rho^n, n = 1, 2, 3, \dots, K \tag{12}$$

M/M/c/K Queueing Model

This model has a single queue served by more than one physician. The number of patients entering the system is limited to *K* because of limited space in the waiting room. The model is an extended version of the *M/M/1/K* model for the motive of reducing waiting time in queueing guises. This model is common to many healthcare facilities with better resources such as labor and equipment. It is the best model to implement in life threatening areas like ED and ICU departments. Many lives are served and few patients are turned back or referred to other hospitals. In addition, ALOS is reduced drastically and server utilization capacity is also reduced paving way to more in-patients to be accommodated in the system. We present steady-state equations (13) and (14) for this system.

$$p_0 = \left\{\sum_{n=0}^{c}\frac{\rho^n}{n!} + \sum_{n=c+1}^{K}\frac{\rho^n}{c!c^{n-c}}\right\}^{-1} \tag{13}$$

$$p_n = \begin{cases}\frac{\rho^n}{n!}p_0, 0 \leq n \leq c \\ \frac{\rho^n}{c^{n-c}m!}p_0, n \geq c\end{cases} \tag{14}$$

RESULTS AND DISCUSSIONS

The summary of operational profile of both the Emergence Department and the Intensive Care Unit Department of a General Hospital is shown in Table 1. We have combined 3 wards of ICU (men's ward, women's ward and children's ward) since it's a small hospital managed by very few physicians. The average length of stay in ED is 6 hours while ALOS in ICU is 5 days. There are few beds in both the ED and ICU since the rooms are currently not large enough but can be expanded when resources permit. The average number of arriving patients in ICU is greater than that of ED (Table 1).

Table-1: Operational Profile for ED and ICU

	ED	ICU
Average Length of Stay (ALOS)	6 hours	5 days
Mean number of arriving patients per month	125.2	137.6
Mean number of patients admitted per month	112.5	135.4
Mean number of patients served per month	129.5	141.4
Available number of beds	10	20
Mean number of patients referred to other hospitals	15.3	9.8

Table 2 shows the operational profile expected from ED assuming an infinity waiting room with two different options that is a single doctor and two serving doctors. When only one doctor is serving patients, we expect that doctor to work flat out with only 33 minutes resting time per day that is 98% of the time is devoted to work. We also expect the queue to be 41 patients with one patient under treatment at any given time. The probability that an arriving patient will be served immediately or has to wait is 0.0233 and 0.945 respectively. In addition, each patient is expected to spend 10 days in ED before being transferred to other departments for hospitalization and the expected delay or ALOS is 10 days with probability of delay of one patient being 0.1. This probability increases as the number of patients increase. The second option of introducing two doctors serving the queue reduces the doctor-utilization- capacity to 66% percent, this means each doctor can rest for 8 hours per day. The expected number of patients in the queue is 0.306 and the probabilities that an arriving patient will be served immediately and has to wait are 0.3438 and 0.3205 respectively. Additionally, the average time a patient is expected to spend waiting in queue and in ED is 1 hour 42 minutes and 7.3 hours respectively with expected delay of 7.3 hours per patient associated with probability of delay of 0.3205 (Table 2).

Table-2: Results of ED operations with infinite waiting room

Measure of Performance	c=1	c=2
ρ	0.9764	0.6562
P_0	0.0233	0.3438
P_1	0.0227	0.3358
$P(n \geq 2)$	0.9540	0.3205
L_q	41.02	0.3060
L	43.0	1.2823
W_q	9.7674	0.07284
W	10 days	0.3054
$P\{W_q > 0\}$	0.9767	0.3205
$P\{W_q > 1\}$	0.8836	0.0167
$P\{W_q > t\}$	$0.9767e^{-0.1t}$	$0.3205e^{-3t}$
$P\{W > t\}$	$e^{-0.1t}$	$1.35e^{-4.3t} - 0.35e^{-5.8t}$
$E(D)$	10.0	0.3054
P_d	$1 - e^{-0.1t}$	0.3205

Table 3 shows the operational profile expected from ICU assuming an infinity waiting room with two different options that is a single doctor and two serving doctors. When considering only one doctor in the ICU department, we expect the doctor to work tirelessly with only 30 minutes resting time per day. We are also expecting 45 patients waiting for treatment at any given time with one patient under treatment. The probability that an arriving patient will be served immediately or has to wait is 0.02123 and 0.958 respectively. Moreover, each patient is expected to spend at least 10 days in ICU before being discharged or transferred to other departments for hospitalization. In addition, the expected ALOS is 10 days with probability of delay of 0.1 per patient. This probability increases with an increase in the number of patients received. The second option (Table 3) requires the service of two doctors serving patients. The number of patients in the waiting line reduces to 0.308 with each doctor spending 66% of his/her time serving patients. That is 8 hours per day. The probability that an arriving patient will be served immediately or has to wait is 0.3428 and 0.3216 respectively. Furthermore, the average time a patient is expected to spend waiting in queue is 1 hour 36 minutes with expected length of 5 hours of hospitalization.

Table-3: Results of the ICU operations with infinity waiting room

Measure of Performance	c=1	c=2
ρ	0.9788	0.6572
P_0	0.02123	0.3428
P_1	0.02078	0.3356
$P(n \geq 2)$	0.9580	0.3216
L_q	45.1	0.3082
L	46.1	1.2870
W_q	9.788	0.0669
W	10	0.2792
$P\{W_q > 0\}$	0.9788	0.3216
$P\{W_q > 1\}$	0.905	0.0128
$P\{W_q > t\}$	$0.9788e^{-0.1t}$	$0.3216e^{-3.2t}$
$P\{W > t\}$	$e^{-0.1t}$	$1.35e^{-4.7t} - 0.35e^{-6.3t}$
$E(D)$	10.0	0.2792
P_d	$1 - e^{-0.1t}$	0.3216

Table 4 shows results of the operations in ED with limited waiting space. The maximum number of patients allowed in the department is only 10. An additional arrival is referred to other hospitals which would be free. We have considered two options. The first option is when we have only one doctor and the second option is when two doctors are hired. Under a single doctor, we expect to lose 3.86 patients per unit time because of limited waiting space. Again, the probability that an arriving patient will immediately get a doctor free is 0.102 and the probability that an arriving patient will get the doctor busy and has to wait is 0.798. Additionally, we expect to have an average of 3.87 patients waiting in the queue with one patient under treatment. Each patient is expected to spend an average of a day in ED before transferred to other departments for further treatment. On the other hand, two doctors in ED (Table 4) reduce the average number of waiting patients in queue and in the system to 0.3 and 1.28 respectively. Moreover, a patient is expected to

spend an average of 7.3 hours in the ED and that is also the ALOS per patient.

Table-4: Steady-State results of ED operations with finite waiting room of 10 per unit time

Measure of Performance	c=1	c=2
ρ	0.9767	0.6561
P_0	0.102	0.3439
P_1	0.09961	0.3359
λ_{eff}	3.8615	4.1978
L_q	3.8670	0.3010
L	4.7650	1.2773
W_q	1.0	0.0717
W	1.234	0.3042
$E(D)$	1.234	0.3042

Table 5 shows results of the operations in ICU with limited waiting space. The maximum number of patients allowed in this department is 20. An additional arrival is referred to other hospitals which are free. In the first option we consider only one doctor and secondly, we consider two doctors to attend to patients. When having a single doctor, we expect to lose 4 patients per unit time because of limited waiting space. Again, the probability that an arriving patient immediately gets a doctor is 0.057 and the probability that an arriving patient gets the doctor busy and has to wait is 0.8867. Furthermore, we expect to have an average of 8.3 patients waiting in the queue with one being hospitalized. Each patient is expected to spend an average of 1.8 hours waiting for the doctor. On the hand, two doctors in ICU will reduce the average number of waiting patients to 0.31 at any given time and each patient is expected to spend an average of 1.6 hours waiting for the doctor who is free to receive their first hospitalization.

Table-5: Steady-state results of the ICU operations with finite waiting room of 20 per unit time

Measure of Performance	c=1	c=2
ρ	0.9809	0.6581
P_0	0.0574	0.3419
P_1	0.0563	0.3354
λ_{eff}	4.4303	4.61
L_q	8.3504	0.3106
L	9.293	1.2915
W_q	1.8848	0.06738
W	2.0976	0.2801
$E(D)$	2.0976	0.2801

CONCLUSION

In healthcare sector, quality services come with a compromise of devoting more resources; labor, waiting space, efficient laboratory equipment and others. Good management also has an important role to play as far as good services are concerned. Few workforce result in prolonged and sluggish queues which are life threatening especially to accident ill patients. We have seen from results above that when there is only one doctor to serve patients in either ED or ICU department, waiting queues are not manageable and this may cause many deaths and long suffering of dear patients. Again few waiting space, (in this regard we are referring to beds) is another crucial variable to take into consideration as some of the critical patients may fail to be accommodated. Referring such patients to other free hospitals could result in long anguish and deaths. Devoting many doctors to each department reduces waiting queues and ALOS dramatically making a system a workable environment.

The general hospital in concern is a moderately small hospital where its operations in ED and ICU departments need immediate attention especially on the medicinal capacity, labor force and waiting space. It is of concern to note that more accidents are occurring along the nearby highway and still few patients have access to the services because of few doctors and limited space at these sister departments. If the $M/M/1$ is adopted in both departments, the number of beds which would be required totals to 43 in ED and 230 in ICU which is a sizeable number not afforded at this hospital. While an additional of one doctor in each department requires 8 beds in ED and 24 beds in ICU. Very few patients would be turned away since the system is stable. Therefore, we recommend to the authorities to implement $M/M/c/K$ model in both department where c and K are determined by the availability and variability of resources per given time. The important advantage of this queueing model is that it reduces the waiting time and the length of stay per patient in the system significantly.

APPENDIX

Notations and symbols used in this research.

n - The number of patients in ED or ICU departments.
P_0 - The steady – state probability that the physician or doctor is free.
$P_n(t)$ – The time dependent probability of exactly *n* patients being in the system at time *t*, given that the system started at time zero.
P_n– The steady-state probability of exactly n patients in the system.
λ – The number of patients arriving per time unit or the mean arrival rate.
μ - The number of patients served per unit time or the mean service rate per busy server.
c - The number of doctors or physicians to attend to patients.
K – The permissible number of patients in the system due to limited space in the waiting room.
$\rho = \frac{\lambda}{\mu}$ – Capacity utilization of the doctors or physicians.
W – The mean waiting time per patient in the system.
W_q– The mean waiting time per patient in the queue.
$L = E\{n\}$ – The mean number of patients in the system.
$L_q = E\{n_q\}$ – The mean number of patients in the queue.
$d = e\{D\}$ – Expected delay per patient.
$F_d(x) = P(D < x), x \geq 0$, - The probability distribution of patient delay.
P_d – The probability of delay per patient.
M – The Markovian arrival or service distribution.

ACKNOWLEDGEMENTS

We would to thank Great Zimbabwe University for encouraging us to work on this article. We extend our special thanks to the management at the general hospital for allowing us to collect sensitive data regarding the overall service at the hospital.

REFERENCES

1. Olorunsola SA, Adeleke RA, Ogunlade TO. Queueing Analysis of Patient Flow in Hospital. Journal of Mathematics. 2014; 10(4) 47-53.
2. Gupta D. Queueing Models for Healthcare Operations; Handbook of Healthcare Operations Management: Denton. B., (2013). Springer. 2013, NY.
3. Brockmeyer E, Halstrom HL, Jensen A. The life and works of A. K. Erlang. Transactions of the Danish Academy of Technical Science 2. 1948
4. Green L, Yankovic N. Identifying Good Nursing Levels: A Queueing Approach. Operations Research. 2011; 59, 942-955.
5. Hall RW. Patient Flow; The new queueing theory for healthcare, OR/MS Today, 2006.
6. Green LV. How many Hospital Beds? Inquiry. 2002; 37(4), 400-412.
7. Armony M, Israelity S, Mandelbaumz A, Marmorx YN, Tseytlin Y, Yom-Tovk GB. On Patient Flow in Hospitals: Data based Queueing Science Perspective. Stochastic Systems. 2015; 50 pages. Available on www.stern.nyu.edu/om/faculty/armony/research/Patient%20flow%20main2.pdf
8. Wang JY. Queueing Theory Chapter 17. NCTU Operations Research II Spring. 2009.
9. Green LV, Giulio J, Green R, Soares J. Using queueing theory to increase the effectiveness of physician staffing in the emergency department, Academic Emergency Medicine. 2005. Unpublished work.
10. Denton BT. Handbook of Healthcare Operations Management: Methods and Applications. Springer. 2013.
11. Green L. Capacity Planning and Management in Hospitals. In Operations Research and Health Care: A Handbook of Methods and Applications. Kluwer Academic Publishers. London. 2004; 14-41.
12. Hall RW. Patient Flow: Reducing Delay in Healthcare Delivery, 2nd Edition. Springer. 2013.
13. Cooper AB, Litvak E, Long MC, McManus ML. Emergency Department Diversion: Causes and Solutions. Acad Emerg Medic. 2001; 8 1108-1110.
14. De Bruin AM, Van Rossum AC, Visser MC, Koole GM. Modeling the Emergency Cardiac In-Patient Flow: An Application of Queueing Theory. Health Care Manag Sci. 2007; 10 125-137.
15. Hall R, Belson D, Murali P, Dessouky M. Modeling patient flows through the health care system, in Patient Flow: Reducing Delay in Healthcare Delivery. In Hall, R. W ed (Hall, R. W. ed., pp. 1-44). New York: Springer. 2006.
16. De Bruin AM, Bekker R, Van Zanten L, Koole GM. Dimensioning Hospital Wards using the Erlang Loss Model. Ann Operations Res. 2009; 178 23-43.
17. Green LV, Kolesar PJ, Whitt W. Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. Prod Oper Manag. 2007; 16 13-39.
18. Green LV. Queueing analysis in healthcare, in Patient Flow: Reducing Delay in Healthcare Deliver. Springer. 2006a;

New York.
19. Worthington DJ. Queueing models for hospital waiting lists. J Oper Res Soc. 1987; 38: 413-422.
20. Arnoud M, Rossum AC, Visser MC, Koole GM. (2006). Bottleneck Analysis of Emergency Cardiac In-patient Flow in a University: An Application of Queueing Theory. Invest Med. 2006; 28(6): 316-317
21. Cooper JK, Corcoran TM. Estimating Bed Needs by Means of Queueing Theory. New England Journal of Medicare. 1974; 344: 404–405.
22. McManus MC, Long MC, Cooper AB, Litvak E. Queueing Theory Accurately Models the Need for Critical Care Resources. Anesthesiology. 2004; 100(5),1271-1276.
23. Koizumi N, Kuno E, Smith ET. A Queueing Network model with blocking: Analysis of Congested Patient Flow in Mental Health Systems. Health Care Management Science. 2005; 8(1):49-60.
24. Cochran KJ, Bharti A. A Multi-stage Stochastic Methodology for whole Hospital Bed Planning Under Peak Loading. Int J Industr and Sys Eng. 2006; 1(1/2): 8-35
25. Afrane S, Appah A. Queueing Theory and the management of Waiting – time in hospitals: A case of Anglo Gold Ashanti Hospital in Gana. Inter J Acad Res Bus Soc Sci. 2014; 4(2). 34 - 44.
26. Kim S, Horowitz I, Young K, Buckly T. A Flexible Bed Allocation and Performance in the Intensive Care Unit. Journal of Operations Management. 2000; 18: 427–443.
27. Obamiro J. Application of Queueing Model in Determining the Optimum number of Service Facility in Nigerian Hospitals, M Sc. Project submitted to Department of Business Administration, University of Ilorin. 2003.
28. Cooper RB. Introduction to Queueing Theory. Macmillan. 1981;New York
29. Zuckerman M. Introduction to Queueing Theory and Stochastic Teletraffic Models. Unpublished Book, 2015. City University of Hong Kong. 2015.