© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715232

Application of Queuing Theory in The Optimization of Waiting Time and Cost in a Teaching Hospital

AKANIHU, C. N.[1], ALI, H.[2], SAKYA M. R.[3]
[1,2]*Department of Mathematics, University of Jos, Jos, Nigeria.*
[3]*Plateau State Polytechnic Barkin Ladi, Jos, Nigeria.*

Abstract- ***The health care system constitutes an essential part of the service sector. Over the years, hospitals have become increasing successful in deploying medical and technical innovations in order to deliver more effective clinical treatment. However, they are still confronted with pressure, delay, congestion and inefficiency presenting ground for research in numerous scientific fields. Providing too much service capacity to operate a system involves excessive cost, but not providing enough service capacity results in long waiting time and cost. Primary source of data collection was adopted using counting process and descriptive observation at the eye clinic of Jos University Teaching Hospital (JUTH). The statistical model was multi-server queuing model and First come, First served queuing discipline. The data was analysed using TORA optimization software as well as using descriptive analysis. The results revealed that the average number of patients in the queue, average number of patients in the system, expected time a patient spent in the queue, expected time a patient spent in the system and the utilization factor could be reduced at an optimal server level and at a minimum expected total cost as against their present server level with maximum expected total cost which include waiting and service costs. Therefore, this call for serious attention in order to improve the overall patient care and satisfaction.***

Index Terms- ***Queue, Hospital, Satisfaction, Waiting Time, Cost.***

I. INTRODUCTION

Queue in facilities that provide essential services is inevitable in the day-to-day activates of life and it is a typical situation that occurs in every-day life. Long waiting times in hospitals have become a major challenge in many healthcare facilities. Long waiting times in hospitals can result to patient dissatisfaction, increased working costs, overcrowding, and reduced efficiency in the dischargse healthcare services. In a nut shell, queue is formed when demand for service exceeds supply. Waiting time rely on the number of customers in the queuing system, the number of servers attending to the queue and the amount of service time for each customer varies tremendously as the case may be [11]. In health sector the adverse effect of queuing in relation to the time spent in a queue for patients to access clinical treatment is increasingly becoming a huge source of concern for modern society that currently exposed to great significant development in technological advancement and the danger of keeping patient waiting could become a cost to them [12]. The time wasted on the queue would have been effectively utilized else-where (opportunity cost). The use of queuing network techniques allows one to capture the stochastic nature of arrivals and service time that is common in health system [13]. Long waiting time in the public health care system have been a major source of concern in most of the countries in the world. Patient flow is a very complex phenomenon because of the random nature of arrival and service mechanisms for the patient [10]. This often requires a systematic approach in future planning and management of facilities. Queuing theory is considered as one of the analytical techniques and accepted as valuable tool to be used in modelling and analysing the arrival and service time for patient coming to a health care system [8].

Queue theory also known as waiting time was first analysed by Agner Krarup Erlang a Danish engineer, mathematician and statistician in 1913 in the context of telephone facilities as he was experimenting with the fluctuating demand for telephone facilities and its effect on automobiles dialling equipment at the Copenhagen telephone system [9]. Since the World War II, this theory has been applied to many business and human services field [13]. Literature on queuing indicates that waiting time in line or queue causes

inconveniences from economic cost to individuals and organizations.

Service pressure, delay, congestion and inefficiency are mostly caused or experience due to high demand for a service and the capacity available to meet the current demand is limited [7]. The variability and interaction between the arrival and service process make the dynamic of service systems often more complex and as a result the prediction of congestion levels becomes difficult [4]. In view of this, there is a need to apply queuing theory model to determine the capacity needed to achieve some levels of performance and improve service standards.

In general, customer's satisfaction is multi-factorial and is considered as part of overall customer behaviour model. Customer evolves over time and is influenced by many factors. Several key factors that greatly influence satisfaction include customer's expectations, attitudes and intention about the service provided. Customer's satisfaction has been defined as the difference between the customer's perceptions of the experience and their expectations, which is many times based on past experience. Although it is possible to manage and decrease actual waiting time and to some extent to manage customer's expectations about customer's satisfaction [5].

II. MATERIALS AND METHODS

Modelling patient's satisfaction using queuing theory is an approach or technique used to examine the possible causes of delay, pressure, inefficiency and congestion there by suggesting possible ways to decongest the queue.

The method of data collection for this research was a primary source using the counting process, direct observation and interviews. The data were collected for four weeks at the eye clinic of Jos University Teaching Hospital (JUTH) and the methods employed during the data collection were counting process, direct observation and personal interview from the staff and patients at the eye clinic.

The following assumptions were made with regard to the data.

i. It is assumed that the arrival time follows a Poisson distribution.

ii. Time of inter-arrival of patients is independent which follows an exponential distribution.

iii. The service time follows an exponential distribution.

iv. It is assumed that patients were served on First come, First serve queuing discipline.

v. It is assumed that the system capacity is infinite.

vi. It is also assumed that the population capacity is infinite

These are extracts from [3].

2.1 Arrival processes

We denote t_i the time at which the i^{th} patient arrives and assume that T's are independent continuous random variables projected from the random variable A. The underlying assumptions that each inter-arrival times are controlled by the same random variables implies that the distribution of arrivals remains independent at different times [6]. This is a stationary inter-arrivals time assumption and it is worth noting that the stationary inter-arrival times are often unrealistic but the approximation can be achieved by breaking the length of time into segments such that a negative inter-arrival time will be impossible to achieve hence,

$P\left(A \le C\right) = \int_0^c a\left(t\right)dt$ and

$P\left(A > C\right) = \int_0^\infty a\left(t\right)dt$, we define $1/\lambda$ to be the mean or inter-arrival. We also define λ to be the arrival rate. It is important to choose A such that the problem remains computationally tractable. The most common choice of A is through the exponential distribution with parameter λ and a density function

$a\left(t\right) = \lambda \ell^{-\lambda t}$,

where $\mathrm{E}(A) = 1/\lambda$ and $Var(A) = 1/\lambda^2$.

Furthermore, as stated by Gross et al., (2018), the arrival process is a stochastic process that describes the sequence of times at which patients arrives at a service facility.

2.2 Service processes

We assume that the service times for different patients are independent random variables and that each patient's service time is controlled by a random variable having a density function $S(t)$. We let $1/\lambda$ to be the mean service time for patients. Generally, service time may not be consistent with the exponential distribution memory-less property and it is by this reason that we assume $S(t)$ is an Erlang distribution with shape parameter K and rate parameter $K\lambda$. In certain cases, the inter-arrival or service times may be modelled with zero variance and under such cases both the inter-arrival and service time are considered deterministic. As an example, if the inter-arrival times are deterministic, then each inter-arrival time will be exactly $1/\lambda$ and similarly for the service times it will also remain the same [2].

As highlighted by [9], the service process is the sequence of times at which customers are served, and is typically characterized by a service time exponential distribution.

2.3 Queuing Models

In specifying a queuing model, we must make some necessary assumptions about the probabilistic nature of the arrival and service processes. The most common assumption to make about the arrivals is that they follow a Poisson process. This results from the fact that the number of arrivals at any given point has a Poisson distribution. So, if $N(t)$ is the number of arrivals during a time period of duration t and N(t) has a Poisson distribution, then

$$N(t) = \frac{(\lambda t)^k \ell^{-\lambda t}}{k!} \quad (1)$$

λ is called the rate and it is the expected number of arrivals per unit of time. Another way to characterize the Poisson is that the time between the consecutive arrivals known as the inter-arrival times (IA) has an exponential distribution. So, if (IA) denotes the inter-arrival times for a Poisson process with rate λ, then

$$P(IA \le t) = 1 - \ell^{-\lambda t} \quad (2)$$

Where $1/\lambda$ is the average time between the arrivals. An important property of the exponential distribution is that it is "memory-less". This implies that the time of the next arrival is independent of when the last arrival occurred. It can be shown analytically that if patients arrive independently from one another, then the arrival process is also a Poisson process and it is for this reason that the Poisson process is considered the most random arrival process. In determining this fact in a specific service system, it is useful to consider the following three properties.

i Patients arrive independently.

ii The probability that a patient arrives at any given time is independent of arrival for the next patient.

iii The probability that the patient arrival at any given time is independent of the time.

The assumption of the Poisson process is sufficient when the three properties listed above accounts for a reasonable description of a system. Further tests can be performed to determine the goodness of fit. Such tests are based on the relationship of the standard deviation and the mean for the two distributions involved in the Poisson process. Since the Poisson distribution is characterized by its standard deviation being equal to its mean, we can then look at the inter-arrival times and compute the ratio of the standard deviation to the mean and check if it is closer to one. [7] stated that a queuing model is a stochastic model

that describes the behaviour of a queuing system, including the arrival process, service process and queuing discipline, and it is used to evaluate system performance metrics or measures such as waiting times and queue lengths.

2.4 Cost analysis model

Two opposing costs must be considered in order to evaluate and determine the cost and number of servers in the system. They are: service costs and waiting time costs of patients. The analysis of these costs helps the management to make a trade-off between the increased costs of providing better service by the specialists and the decreased waiting time costs of patients derived from providing that service. Therefore, expected service costs, expected waiting cost and the expected total cost is given below as;

$\mathrm{E}Sc = CCs$, this is the expected service cost where C is the number of servers and Cs is the service cost of each server.

$\mathrm{E}Wc = (\lambda Ws)Cw$, this is the expected waiting cost where λ is the number of arrivals, Ws is the expected time a patient spends in the system and Cw is the opportunity cost of waiting by patients.

$\mathrm{E}Tc = CCs + (\lambda Ws)Cw$, this is the expected total cost of the system

2.5 Model specification

This research adopted the multi-server queuing model which is stated as:

$M \mid M \mid C : FCFS \mid \infty \mid \infty$

Where,

M = Markovian Poisson arrival rate

M = Markovian exponential service rate

C = Number of servers

$FCFS$ = First come, first serve queuing discipline

∞ = Infinite system capacity

∞ = Infinite population capacity

2.6 Characterization

A queuing system may be described as a logical phenomenon where patients arrive according to an arrival process to be served by a service facility according to a service process. Each service facility may contain one or more servers and it is generally assumed that each server can only service one patient at a time. If all servers are busy, then then the patient is then forced to queue for a service. If a server becomes free again, then the next patient is then picked from the queue according to the rules given by the queuing discipline. During service a patient might run through one or more stages of service, before departing from the system. In queuing theory, models are used to describe the characteristics of a queuing system. Some of the most commonly considered characteristics are discussed below:

L_q is the queue length which is the average number of patients in the queue waiting to get a service. Large queues may indicate poor server performances while small queue may indicate too much server capacity. It is computed with the formula given by

$$L_q = \frac{\lambda\mu\left(\lambda/\mu\right)^C}{(C-1)!(C\mu-\lambda)^2}\rho_0 - \lambda/\mu \qquad (3)$$

OR $L_q = L_s - \lambda/\mu$

L_s is the system length which is the average number of patients in the queue within the system, that is, those waiting and those that are being served. Large values of these statistics imply congestion and possibly patient's dissatisfaction and potential need for greater service capacity.

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715232

Computational formula is given as:

$$L_s = \frac{\lambda\mu\left(\lambda/\mu\right)^C}{(C-1)!(C\mu-\lambda)^2}\rho_0 + \lambda/\mu \qquad (4)$$

OR $L_s = L_q + \lambda/\mu$

W_q is the waiting in the queue explains the average time that a patient has to wait or spend2 in the queue to get service. Long waiting times are directly related to the patient's dissatisfaction and potential loss of future revenues while very small waiting times may also indicate too much service capacity.

$$W_q = \frac{\mu\left(\lambda/\mu\right)^C}{(C-1)!(C\mu-\lambda)^2}\rho_0 \qquad (5)$$

Or $W_q = W_s - \frac{1}{\mu}$

W_s is total time in the system gives the average time that a patient spent in the system, from the instant of joining the queue to service completion. Large values of this statistic is an indication of the need to make adjustment in the service capacity.

$$W_q = \frac{\mu\left(\lambda/\mu\right)^C}{(C-1)!(C\mu-\lambda)^2}\rho_0 + \frac{1}{\mu} \qquad (6)$$

ρ_0 is the service idle time that explains the relative frequency within which the service system is idling and this idle time is directly related to cost. However, reducing idle time may have adverse effects on the other characteristics mentioned.

$$\rho_0 = \left\{\left[\sum_{i=1}^{C-1}\frac{1}{t!}\left(\frac{\lambda}{\mu}\right)^i\right] + \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C \frac{1}{\left(1-\lambda/C\mu\right)}\right\}^{-1} \qquad (7)$$

III. RESULTS AND DISCUSSIONS

Table 1 shows the number of patient's arrival per hour (λ), service time per hour (μ) and the number of servers in the eye clinic Jos University Teaching Hospital (JUTH). Table 2 shows the performance measures of multi-server queuing model surveyed by the researcher. Table 3 displays the performance measures of the multi-server queuing model and the optimality cost. Table 4 shows the summary analysis of the multi-server queuing model and table 5 shows the correlation between the utilization factor, L_q, W_q, L_s and W_s. It should be noted that we use the average values for λ and μ for different scenarios in this research.

Tables 1: Input parameters

Scenario	C	λ	μ	∞
1	8	17	4	Infinity
2	9	14	6	Infinity
3	10	20	12	Infinity
4	11	13	4	Infinity
5	12	11	3	Infinity

Table 2: Performance measures of M/ M/ C: FCFS / ∞ / ∞

Scenario	C	λ	μ	ρ_0	L_s	L_q	W_s	W_q
1	8	17	4	0.01408	4.33989	0.08989	0.25529	0.00529
2	9	14	6	0.09697	2.33359	0.00026	0.16669	0.00002
3	10	20	12	0.18888	1.66667	0.00000	0.08333	0.00000
4	11	13	4	0.03877	3.25025	0.00025	0.25002	0.00002
5	12	11	3	0.02556	3.66687	0.00020	0.33335	0.00002

Table 3: Performance measures of M/ M/ C: FCFS / ∞ / ∞ and optimality cost

Scenario	C	λ	W_s	λW_s	C_s (₦)	C_w (₦)	CC_s (₦)	(λWs)Cw(₦)	ETc (₦)
1	8	17	0.25529	4.339933	3,225.02	5,142.61	25,800.16	22,318.57	48,118.74
2	9	14	0.166669	2.333366	3,225.02	5,142.61	29,025.18	12,001.10	41,026.28
3	10	20	0.083333	1.666660	3,225.02	5,142.61	32,250.20	85,70 0.67	40,820.88
4	11	13	0.250002	3.250026	3,225.02	5,142.61	35,475.22	16,714.82	52,190.04
5	12	11	0.333335	3.666685	3,225.02	5,142.61	38,700.24	18,857.18	57,557.42

Table 4: summary analysis of the M/ M/ C: FCFS / ∞ / ∞ and optimality cost

Number of servers (C)	8	9	10	11	12
Arrival rate (λ)	17	14	20	13	11
Service rate (μ)	4	6	12	4	3
Utilization factor (ρ)	0.53125	0.25926	0.16667	0.29545	0.30556
L_s	4.33989	2.33359	1.66667	3.25025	3.66687
L_q	0.08989	0.00026	0.00000	0.00025	0.00020
W_s	0.25529	0.16669	0.08333	0.25002	0.33335
W_q	0.00529	0.00002	0.00000	0.00002	0.00002
ρ_0	0.01408	0.09697	0.18888	0.03877	0.02556
ETC (₦)	48,118.74	41,026.28	40,820.88	52,190.04	57,557.42

3.1 Correlation between utilization factor, L_q, W_q, L_s and W_s

As result shown by using queuing analysis where utilization factor increased or decreased as average number of patients waiting in the queue and average waiting time varies as the case may be. We used Minitab software to measure the strength of relationship between these characteristics of queuing theory.

Table 5: Strength of the relationship between utilization factor, L_q, W_q, L_s and W_s

Pairwise Pearson Correlations

Sample 1	Sample 2	N	Correlation	95% CI for ρ	P-Value
L_s	ρ	5	0.901	(0.090, 0.993)	0.037
L_q	ρ	5	0.914	(0.165, 0.994)	0.030
W_s	ρ	5	0.568	(-0.630, 0.966)	0.318
W_q	ρ	5	0.915	(0.168, 0.994)	0.030
L_q	Ls	5	0.679	(-0.506, 0.976)	0.207
W_s	Ls	5	0.859	(-0.094, 0.991)	0.062
W_q	Ls	5	0.680	(-0.505, 0.976)	0.206
W_s	L_q	5	0.221	(-0.821, 0.923)	0.720
W_q	L_q	5	1.000	(1.000, 1.000)	0.000

W_q	W_s	5	0.223	(-0.821, 0.923)	0.719

3.2 Graphical Representation of The Queuing Characteristics

Graph figure 1, 2, 3, 4, 5, 6 and 7 displayed the line plots of the utilization factor (ρ) against the average waiting time in the system (W_s), number of servers (C) against the probability of the system being idle (ρ_0), expected service cost (CCs) against the number of servers (C), expected waiting cost against the number of servers (C), expected total cost against the number of servers (C), utilization factor (ρ) against the number of servers (C) and average number of patients in the system (L_s) against probability of the system being idle (ρ_0).

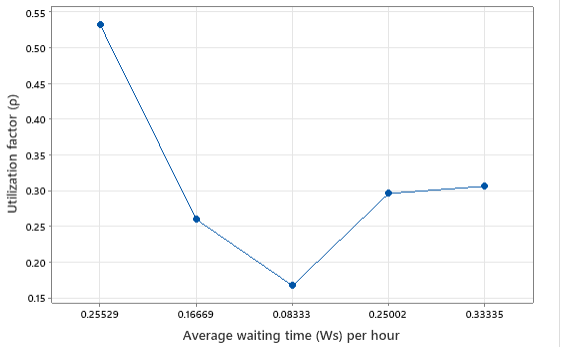

fig.1: Utilization factor(ρ) against the average waiting time in the system(Ws)

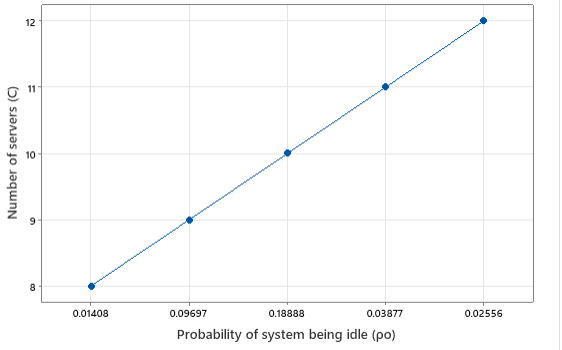

fig.2: Number of servers (C) against probability of system being idle (ρo)

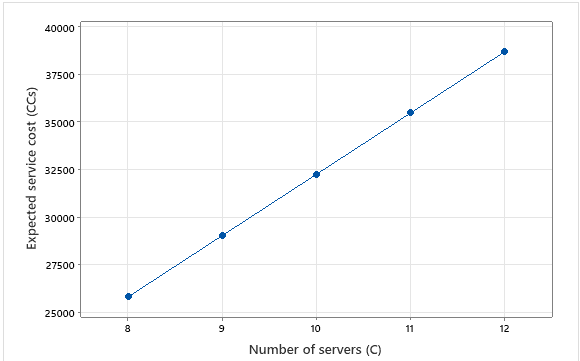

fig.3: Expected service cost (CCs) against the number of servers (C)

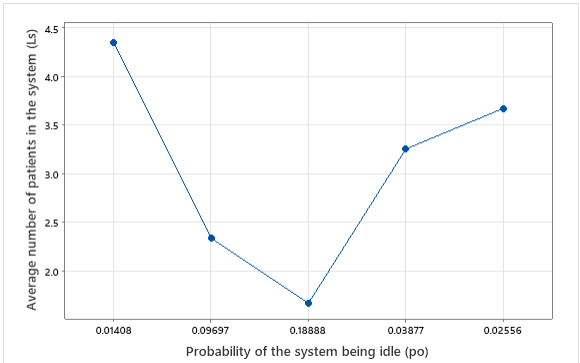

fig.7:Ave. no. of patients in the sys.(Ls) against prob. of sys. being idle(ρo)

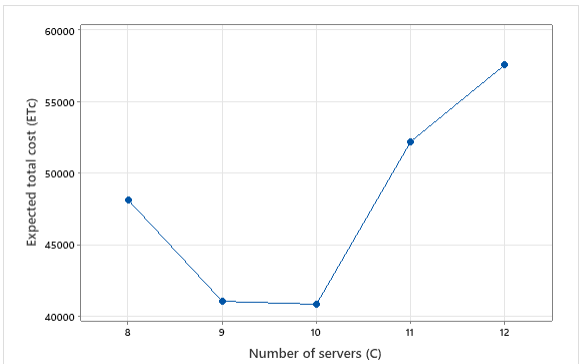

fig.5: Expected total cost (ETc) against the number of servers (C)

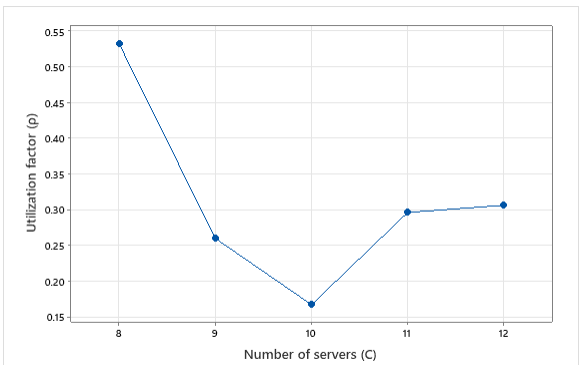

fig.6: Utilization factor (ρ) against the number of servers (C)

© MAR 2026 | IRE Journals | Volume 9 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I9-1715232

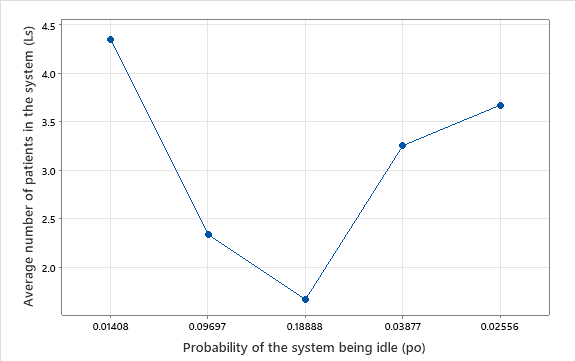

fig.7:Ave. no. of patients in the sys.(Ls) against prob. of sys. being idle(po)

3.2 DISCUSSIONS

Scenario 1: the system performance measures are as follows:

$L_q = 0.08989$, this implies that there are 0.08989 patients on the queue waiting to be served by doctors.

$L_s = 4.33989$, this means that there are 4.33989 patients in the system waiting to get service from the service facility.

$W_s = 0.00529$, this means that patients spent 0.00529 hours on the queue waiting to be attended by doctors.

$W_s = 0.25529$, means that patients spent 0.25529 hours (15 minutes) in the system.

$\rho_0 = 0.01408$, this implies that the probability that the system is idle is 0.01408.

$\rho = 0.53125$, this is the probability that the system is busy.

Scenario 2: the performance measures are as follows:

$L_q = 0.00026$, this is the average number of patients on the queue waiting for service by doctors.

$L_s = 2.33359$, this the average number of patients in the system.

$W_q = 0.00002$, this implies that patients spent 0.00002 hours on the queue.

$W_s = 0.16669$, this means that patients spent 0.16669 hours (10 minutes) in the system.

$\rho_0 = 0.09697$, the probability that the system is less busy or idle is 0.09697.

$\rho = 0.25926$, The probability that the system is overworked or busy is 0.25926.

Scenario 3: the performance measures seems to be appreciable than scenario 1 and 2.

$L_q = 0.00000$, this means that the average number of patients waiting on the queue to get service is 0.00000.

$L_s = 1.66667$, this is the expected number of patients in the system waiting for service by the doctor.

$W_q = 0.00000$, this is the expected waiting time on the queue by patients to get service by doctors.

$W_s = 0.0833$, this implies that patients spent 0.0833 hours (5 minutes) in the system.

$\rho_0 = 0.18888$, this is the probability that the system is idle.

$\rho = 0.16667$, this is the probability that the system is busy.

From all the scenarios, scenario 3 with a 10-number of doctors is better than 8-number of doctors in terms of performance measures. With regard to cost implication, a 10-numbers of doctors in the system records the minimum or lowest cost of ₦40,820.88 compared to 8-number of doctors that records ₦48,118.74. These costs include both waiting cost and service cost. This is in line with the findings by Bailey (1954), which established that in out-patients and in-patients' clinic, when the number of servers is below a certain threshold, a clinic develops an infinite queue whereas when it is slightly above this threshold, waiting time and queue length are lower. However, one need to be careful of the cost involved in other to achieve all these marginal changes.

Employing more doctors will mean taking on more costs. A good balance between the number of doctors, cost and optimal system performance is crucial for sustainability.

Similarly, adopting the same approach, Singh (2006) looked into minimizing total cost incurred and also minimizing the waiting time by comparing the output of two nurses, three nurses and four nurses by evaluating the performance measures for each of the scenario. In that study, it was found that scenario of three nurses was the optimal solution with optimum trade-off between the two types of costs involved in queuing models.

The analysis depicts that the average queue length, delay, pressure, congestion and long waiting time by patients could be reduced or cut down when the number of doctors are increased at a minimum expected total cost of ₦40,820.88. Furthermore, from the correlation it was found that;

- The correlation between utilization factor (ρ) and the average number of patients in the system (L_s) are statistically significant since the p-value 0.037 is less than 0.05 level of significance. The strength of relationship is 0.901 which shows that they are highly correlated.
- The correlation between utilization factor (ρ) and average number of patients on the queue (L_q) are statistically significant since the p-value 0.030 is less than 0.05 level of significance. The strength of relationship is 0.914 which shows that they are highly correlated.
- The correlation between utilization factor (ρ) and the expected time patients spent in the queue (W_q) are statistically significant since the p-value 0.030 is less than 0.05 level of significance. The strength of relationship is 0.915 which shows that they are highly correlated.
- The correlation between the average number of patients on the queue (L_q) and the average time patients spent in the queue (W_q) are statistically significant since the p-value 0.000 is less than 0.05 level of significance. The strength of relationship is 1.000 which shows that they are perfectly correlated.

REFERENCES

[1] Adele, M. & Barry, S. (2005). Modelling patient's flow in hospital using queuing theory. unpublished manuscript.

[2] Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (2001). Queuing networks and Markov chains: Modelling and performance evaluation with computer science applications. *John wiley and sons*, pp. 209-262

[3] Bunday, B. D. (1996). An Introduction to queuing theory. New York: Halsted press.

[4] Davis, M. & Vollman, T. E. (1990). A frame work for relating time waiting and customer's satisfaction in a service operation. *Journal services marketing* 4(1): 61-69.

[5] Fink, R., & Gillett, J. (2006). Queuing Theory and the Taguchi Loss Function: The cost of customer's dissatisfaction in waiting lines. *International Journal of strategic cost management.* Spring. 3(5): 1-9.

[6] Gross, D., Shortle, J. F., Thompson, J. M., & Harris, C. M. (2018). Fundamentals of queuing theory (5th ed.) Wiley.

[7] Kawanishi, K. (2020). Queuing theory: A modern perspective. Springer.

[8] Kembe, M. M., Onah, E. S & Lorkegh, S. A. (2012). A study of waiting and service cost of multi-server queuing model in specialist hospital. *International Journal of Scientific and Technology Research* 5(2): 2277-8616.

[9] Kleinrock, L. (1975). Queuing systems: Theory. Willey

[10] Medhi, J. (2003). Stochastic models in queuing theory, Amsterdam: Academic press, second edition.

[11] Nosek, A. R., & Wilson, P. J. (2001). Queuing theory and customer's satisfaction: A review of terminology, trends and application to pharmacy practice. *Hospital pharmacy*. 36(3): 275-279.

[12] Somani, S. M., Daniels, C. E., Jermstad, R. L. (1982). Patient's satisfaction with out-patients

pharmacy services. *AM J Hosp. Pharm.* 3(9): 1025-7.

[13] Stewart, W. J., (1946). Probability, Markov Chains, Queuing and Simulation, (1st Edition): The mathematical of performance modelling, Princeton University press.