Earthline Journal of Mathematical Sciences
E-ISSN: 2581-8147
Volume 15, Number 3, 2025, Pages 437-454
https://doi.org/10.34198/ejms.15325.437454

Applications of Queue Models to Enhance Effective Healthcare Delivery in Government Hospitals in Nigeria

Simon A. Ogumeyo[1], Esosa Enoyoze[2], Godspower A. Eriyeva[3], Kelvin O. Iyenoma[4], Festus C. Opone[5] and Solomon A. Uriri[6]

[1] Department of Mathematics, Delta State University of Science and Technology, Ozoro, Nigeria

[2] Department of Mathematics, Edo University, Iyamho, Edo State, Nigeria

[3] Department of Statistics, Delta State University of Science and Technology, Ozoro, Nigeria

[4] Department of Information Technology, Federal Medical Centre, Asaba, Delta State, Nigeria

[5] Department of Statistics, Delta State University of Science and Technology, Ozoro, Nigeria
e-mail: oponef@dsust.edu.ng

[6] Department of Physics, Delta State University of Science and Technology, Ozoro, Nigeria

Abstract

The need to develop queue models which could guide hospital management personnel in making policies which enhance prompt healthcare delivery in order to sustain patients' interest and patronage cannot be over-emphasized. Problem of stampede leading to loss of lives in palliative distribution centers has been very frequent in recent times. Traffic congestion is a common phenomenon that occurs when patients' arrival rate surpasses the rate of service in any service providing facilities such as hospitals and clinics. In this research, we present two queuing models with the aim of applying them to solve problem of long queues in hospitals. In model 1, we apply the queue discipline approach of first-come first-served to derive the distribution probability that govern the numbers of patients' arrival and departure in any given time interval of a single-server queuing system. We observe that the expected inter-arrival and expected treatment times of a patient is a continuous density function similar to that of a renewal process. In model 2, we used the method of Komolgorov linear differential equations for each value of $P_n(T)$ to derive the transient solution. The values of $P_n(T)$ is also examined as time T tends to infinity (∞) and we observed that the queue system can reach its statistical equilibrium state, if $P_n(T)$ tends to a limit P_n, and $E[n]$ is a finite value for the distribution limit. The results from the numerical illustration show that the time spent in the queue, the number of patients, and the line length all increase rapidly as the traffic intensity ρ increases. It is also observed that the queue system to attain a steady-state equilibrium for sufficiently large $\rho < 1$, will take a long time. Our proposed models have advantage over the existing ones in that they contain mathematical formulas which could guide hospital management personnel to make policies which enhance prompt services, sustainability of patients' interest and patronage.

1. Introduction

Queues are waiting lines made up of customers waiting to receive services from servers in places such as banks, hospitals, railway stations, airport, traffic lights, etc. A queuing process is made up of customers who arrive at a service facility, then wait in a queue until they are served and ultimately depart from the facility.

Received: January 28, 2025; Accepted: March 6, 2025; Published: March 21, 2025

2020 Mathematics Subject Classification: 60K25, 60K30.

Keywords and phrases: patients, inter-arrival time, queue, arrival rate, service rate.

*Corresponding author

Queues are formed due to inadequate personnel to attend to customers or poor state of facilities used in serving customers. It is difficult to predict the rate of customers' arrival and departure in a service facility at a given time in a real-life situation. There are cases where customers may decide to wait in queues until they are attended to by service providers, while in some other cases customers cannot afford to wait for service if they perceive it will take a longtime before they are attended to. The behaviors of customers are broadly classified into two categories: patient and impatient customers. Impatient customers are further categorized into three classes in Montazer *et al.* (1996) and Sztrik (2012).

The key components of a queue are customers, servers, patterns of customers' arrival, the order in which they are served, the number of servers and how many customers the facility can accommodate per time, which includes those in service and those in the queue. There are two kinds of arrivals in queuing theory: single and bulk arrivals. Single arrival occurs when one customer comes to a service facility per time, while in a bulk arrival more than one customer arrives at a service facility per time. The length of the intervals is assumed to be independently and uniformly or identically distributed and is liken to a renewal process or a sequence of recurrent events as stated in Cobolat (2020). Zhang *et al.* (2016) states that a steady state in a queuing system refers to a state where the arrival rate and service rate are the same for a long period of time. It also implies to a scenario where the queue length is independent of time.

2. Literature Review

Several queuing models which address the problems of patients waiting in hospitals' queues and staff allocation abound in literature. For examples, Worthington (1987) and Belson (1988) models discuss how patients visiting hospitals can be scheduled in order to reduce waiting time in queues. Similarly, Afrane and Appah (2014) developed a queue model which deals with minimization of waiting time in government hospitals in Ghana by using Anglo Ashanti hospital as a case study. A queuing model to enhance patient-flow and reduce waiting time in queues in health-care centers was developed in Green (2006). The impact of waiting time in queues by patients in a primary healthcare clinic and hospitals is reported in Lohlun (2015) and Ba *et al.* (2017), while models to minimize patients' waiting time in general hospitals with emphasis on how to effectively manage queues by reducing crowd congestion and staff allocation are developed in Paul *et al.* (2021) and Mahala *et al.* (2023).

Mathematical models which incorporate customers' reaction as a result of long queues are reported in Natvig (1975) and Rajadurai (2018). A queue system could be in any of the following state: steady state, transient or idle state. The idle state is an inactive period in which the server is idle because of the absence of customers in the facility to receive service. A transient state of a queue occurs when the arrival rate and service rate of customers are not equal. For example, a single-server queue model's performance in a transient state is discussed in Kurmar *et al.* (1993), Singh *et al.* (2013), Singh *et al.* (2014) and Cobolat (2020). Queue models which adopt Markov probability matrix approach in determining queue performance measures are reported in Havi and Kerner (2007), Economou and Kanta (2008). Single-server queue models which study balking customers' behaviors are reported in Natvig (1975) and were later expanded in Abou-El-Ata *et al*. (1992), and Nithya and Haridass (2016). As stated in Falin and Templeton (1997), retrial queues are queues made up of customers who could not assess the service of a facility but keep coming until they are attended to. The most commonly used queue model is the single queue with unlimited waiting space leading into an identical server S, Jain *et al.* (2014).

Retrial queue models have been applied to several networks, such as telephone systems, computer and telecommunication works in Zhang and Wang (2021). The rationality behind joining or not joining a queue is studied in Hogarth (1987), Hassin (2016), Wang and Zhang (2013), etc. The above authors came to the conclusion that cost structure or reward is a factor which could influence customers' rational in joining or not joining a queue. Bounded rationality of customers in single-server queuing models which consider seen and unseen settings is studied in Huang *et al.* (2013) and Li *et al*. (2016). Lie *et al.* (2017) developed a queue which has many servers with multiple objective functions when using matrix approach. Chai *et al*. (2019), later expanded Lie *et al.* (2017) model by considering a situation where the customer does not have all the relevant information about the quality of the service of the queue system. Huang and Chen (2015), Ren and Huang (2018) and Ogumeyo and Emenonye (2023) remarked that customers tend to join queues based on their previous experiences if they cannot evaluate the quality of a queue system.

Emphasis on fast, efficient and smooth healthcare delivery began to gain attention in the 1990s due to advanced technology. One of the ways to solve the problems of long queues in healthcare centers is to schedule doctors and nurses to match patients' pattern of arrivals, group them and treat them according to the nature of their health challenges (Mahala *et al.,* 2023). According to Belson (1988), poor commitment to duty by hospital staff can cause long queues or delay in healthcare delivery. Wang and Zhang (2021), and Lohlum *et al.* (2015) opined that insufficient human and material resources, lack of proper coordination and management of medical personnel and material resources are the causes of long queues in hospitals where patients have to spend a lot of time before receiving required services. Quality healthcare is closely connected to timely assessment of healthcare. Relevant queue models can be used to reduce waiting time in healthcare centre. Queue models can also be used to estimate capacities of future requirements in terms of human and material resources during natural disasters such as Covid-19. A queue model to estimate capacity levels of waiting spaces in hospitals and other healthcare sectors by making policies which enhance effective allocation of nurses, doctors and hospital beds is reported in Paul *et al.* (2021).

The first-come first serve queue discipline is the most popular rule in queuing systems except in hospitals and other healthcare centers where priority rule in cases of emergencies where patients' lives are at risk (Paul *et al.,* 2021). In emergency cases, a service in progress can be interrupted due to arrival of patients in critical conditions in a hospital. The causes of patients waiting in long queues in healthcare centre are enumerated in Mahala *et al*. (2023). Waiting time and utilization of queuing models in healthcare centers are reported in Sudesh *et al*. (2017). Long queues in hospitals can delay treatment of patients by doctors which ultimately prolong their suffering, or leads to their death Paul *et al.* (2021).

The research aims at:

(i) Identifying shortcomings of existing queue models related to healthcare centers.

(ii) Formulating queue models that could be used to determine the mean arrival rate of patients, mean service rate of patients, the probability of having a specific number of patients in a healthcare system which minimizes the cost of service.

The proposed models are extension of earlier works of Paul *et al.* (2021) and Mahala (2023).

Mahala, *et al.* (2023) presents a queuing model to reduce Out Patient Department waiting time in hospital operations. Their model does not contain mathematical formulas that could be used to determine the optimum values of transient solution, statistical equilibrium and stationary probabilities of a queue system. Likewise, Paul *et al.* (2021) developed a queuing model to optimize patient-waiting time in a public hospital in India lack these essential formulas.

Our proposed models have advantage over the existing ones in that they contain mathematical formulas that could be used to determine the optimum values of transient solution, statistical equilibrium and stationary probabilities of a queue system. This could guide hospital management personnel to make policies which shorten queues due to prompt services and enhance the sustainability of patients' interest and patronage. This in effect, leads to high revenue generation and profit maximization. Another advantage of these two models is the prevention of stampede which could lead to loss of lives such as was experienced in Ibadan, Abuja and Okija (all in Nigeria) during the distribution of palliatives in December, 2024.

3. Problem Definition

The major problem plaguing healthcare centers in Nigeria is inability to access prompt services and improved traffic flow of patients. Research has shown that the problem of patients including pregnant women who wait for hours or days in healthcare centers in order to receive medical services is caused by the following factors:

(i) Lack of commitment to duty by hospital staff (doctors and nurses).

(ii) Insufficient human and material resources.

(iii) Lack of proper coordination and management of medical personnel and material resources.

(iv) Lack of relevant queue models to estimate capacities of future required medical personnel and waiting spaces in hospitals to enhance effective allocation of nurses, doctors and hospital beds.

The approach adopted in handling queue system in a firm can either hinder or enhance customers' patronage. The four factors listed above lead to formation of long queues in hospitals where patients have to spend a lot of time waiting before receiving required services. The consequences of have long queues in healthcare centers lead to the problems which this research aims to address:

(i) Long queues

(ii) Patients' dissatisfaction and low patronage.

(iii) Poor revenue generation.

(iv) Stampede.

4. The Mathematical Model

Model Assumption and Mathematical Notations

The Model assumptions are:

(a) The population consists of n patients seeking treatment in a hospital.

(b) Patients' arrival at the healthcare centre is randomly distributed.

(c) Patients join the queue to register for treatment irrespective of the queue length.

Mathematical Notations

$\emptyset =$ Average number of patients arriving for treatment in a given time interval. (i.e. arrival rate),

$\frac{1}{\emptyset} =$ Mean time between the arrivals of patients,

$l =$ The interval length $[T, T + h]$ of time,

$f(t) =$ Time interval between two successive arrival or departure of patients from the treatment centre,

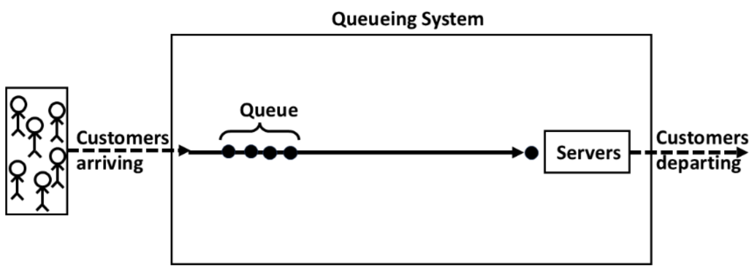
$\psi =$ Average number of patients being treated per time,

$\frac{1}{\psi} =$ Mean time between departures of patients from the treatment centre.

Model 1: Description

In our proposed model, the distribution probability that govern the numbers of patients' arrival and departure in any given interval depends on the length of the interval, rather than the starting point. The model has exponentially distributed inter-arrival time, with parameter $\emptyset$, exponentially distributed service time, with parameter ψ, one server and unlimited capacity. The queue's rule being observed is first-come first-served, where $\emptyset$ denotes the average patients' arrival rate while ψ is the average rate of treatment of patients or service rate. The expected inter-arrival time and expected treatment time to treat one patient are $\frac{1}{\emptyset}$ and $\frac{1}{\psi}$ respectively. The random arrivals of patients in the healthcare centre for treatment indicates that the arrivals are based on an interval length of time, i.e., $[T, T + t]$. While the probability distribution of inter-arrival times describes a continuous density function in which the patient arrival is called renewal process. Taking $f(t)$ as the probability density function (pdf) for the interval time between successive arrivals of two patients at the health centre, $\frac{1}{\emptyset}$ is the mean time between arrival of patients and $\emptyset$ is arrival rate per unit time. Similarly, ψ is the mean number of patients being treated per time (service rate), while $\frac{1}{\psi}$ is the mean time between departures of patients from the healthcare centre.

Figure 1a – Figure 1d show the four major types of queuing system in literature. Figure 1a is a single queue with only one server available to attend to customers. Figure 1b is a single queue with multiple servers in parallel lines while Figure 1c is multiple queues with multiple servers. Figure 1d is a single queue with multiple servers in series. Each of the queuing system has entry point where customers arrive at the facility and exit point where customers who have been served exit the facility. What differentiates one queue from another is the manner in which the customers line up to be served.



(a) Single queue, single server

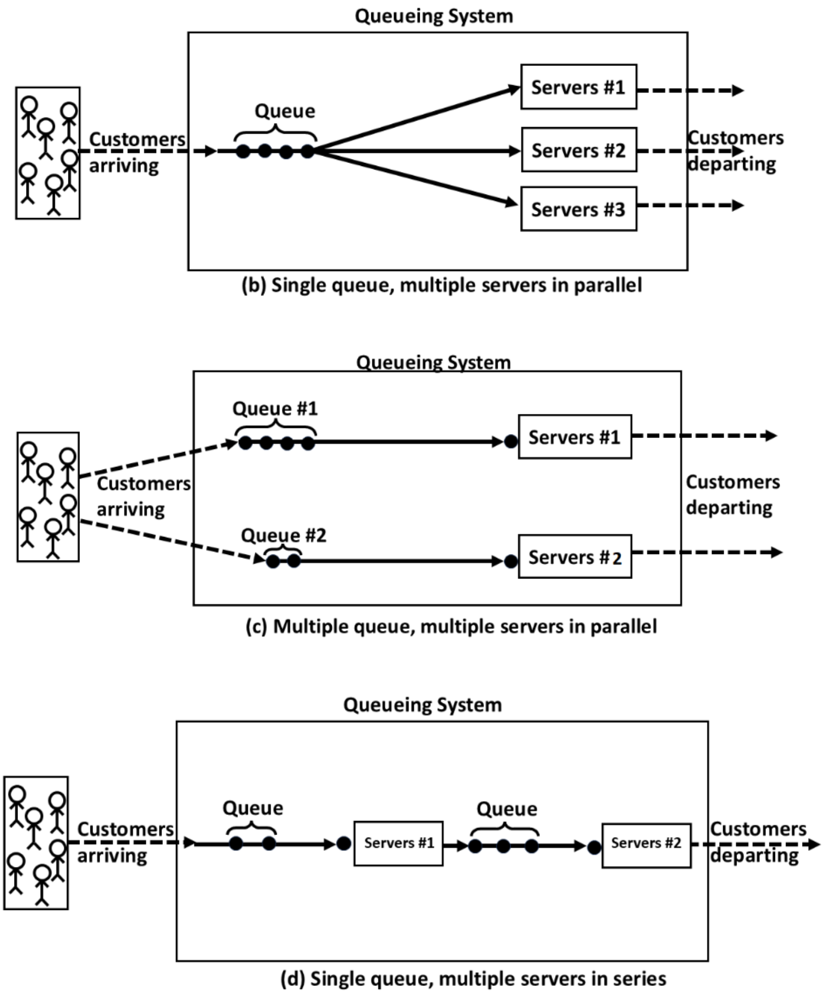


Figure 1: Four types queuing system.

Model 1: Formulation

From the mathematical symbols defined in this section, we can evaluate $\emptyset$ from the density function $f(t)$ by taking the mathematical expection of t. Hence,

$$\int_0^\infty t\, f(t)dt = \frac{1}{\emptyset}. \tag{1}$$

Equation (1) is the meantime arrival of patients. The assumption that patient's arrival is random (that is no fixed pattern) corresponds to Poisson negative exponential distribution. Hence, we have

$$f(t) = \emptyset e^{-\emptyset t}, \quad \text{for } t \geq 0, \tag{2}$$

where $e = 2.7182$, the mean and variance of the distribution are $\frac{1}{\emptyset}$ and $\frac{1}{\emptyset^2}$, respectively.

The probability that no patient was treated in the healthcare centre in the interval $[0,T]$ is equivalent to the probability that the first patient arrived for treatment after time T. This can be stated mathematically as

$$P[t \geq T] = \int_0^\infty \emptyset \ell^{-\emptyset} dt = \ell^{-\emptyset T}. \tag{3}$$

The conditional probability that no patient came for treatment in the interval $[0, T+1]$ given that no patient came for treatment in the time interval $(0, T)$ can be defined as stated in Equation (4).

$$\frac{P[t \geq T+1]}{P[t \geq T]} = \frac{\ell^{-\emptyset(T+1)}}{\ell^{-\emptyset T}} = P[t \geq l]. \tag{4}$$

We observe that Equation (4) depends only on the value of l and it implies that the probability of no treatment for a specified time interval $(T, T+l)$ will be the same even if there was treatment in the time interval $(0, T)$ or at time T and thereby 'renews' the arrival process. If there are n patients who arrive for treatment in $(0, T)$ interval, then the n patients' times of arrival are identically and independently distributed in $(0, T)$ time interval since the interval times will be exponentially distributed. By applying the assumption of exponential interval times $\ell^{\emptyset l}$, we have its Taylor's power series expansion as stated in Equation (5).

$$P\begin{bmatrix}\text{no arrival in}\\ \text{any interval}\\ \text{of lenght } l\end{bmatrix} = \ell^{-\emptyset l} = 1 - \emptyset l + \frac{(-\emptyset h)^2}{2!} + \frac{(-\emptyset h)^3}{3!}. \tag{5}$$

The term $1 - \emptyset l$ in Equation (5) has a larger value compare to the remaining terms. Hence, $1 - \emptyset l$ can be used as an approximate probability in equation (5) when l is very small. Therefore, Equation (5) can be rewritten as

$$P\begin{bmatrix}\text{no arrival in}\\ \text{any interval}\\ \text{of lenght } l\end{bmatrix} = 1 - \emptyset l. \tag{6}$$

Following the same line of argument, the probability of having one patient arriving in the health centre is

$$P\begin{bmatrix}\text{one patient arrival}\\ \text{in an interval}\\ \text{of lenght } l\end{bmatrix} = \emptyset l. \tag{7}$$

Suppose the density function for patient's inter-arrival is exponentially distributed in Equation (2), it therefore means that the total time x for density function in n consecutive patients' arrival will assume a gamma distribution, (See Ogumeyo and Emonenye, 2023). Hence, Equation (2) could be stated as:

$$G(x) = \emptyset(\emptyset x)^{n-1}\ell^{-\emptyset x}, \quad \text{for } x \geq 0, \tag{8}$$

where x is the total of n independent patients obtained from the density function in Equation (2). Hence the total interval for n consecutive patients' arrival can be stated as:

$$P\begin{bmatrix}n \text{ arrivals in a}\\ \text{consecutive interval}\\ \text{length} \leq T\end{bmatrix} = \int_0^T G(x)dx = 1 - \sum_{j=0}^{n-1} (\emptyset T)^j \ell^{-\emptyset T}. \tag{9}$$

Equation (9) can be proved by repeated application of integration by parts technique. Since the probability distribution of patients' arrival in the healthcare centre in any interval time T has Poisson's distribution properties, Equation (9) can be rewritten as.

$$P\begin{bmatrix}n \text{ patients arrival}\\ \text{in any interval}\\ \text{time } T\end{bmatrix} = \frac{(\emptyset T)^n \ell^{-\emptyset T}}{n!}, \quad \text{for } n \geq 0. \tag{10}$$

Since Equation (10) is a Poisson distribution, the expectation and variance can be stated as

$$E[n/T] = \emptyset T \ \text{ and } \ Var\,[n/T] = \emptyset T. \tag{11}$$

It therefore follows from equation (9) and (10) that

$$P\begin{bmatrix}\text{total interval for}\\ n\text{ consecutive arrivals}\\ \leq T\end{bmatrix} = P\begin{bmatrix}n\text{ number of patients}\\ \text{who arrive in any}\\ \text{interval } T \geq n\end{bmatrix}. \tag{12}$$

Similarly, a probability distribution of patient treatment time can be derived. Suppose $G(t)$ is the density function of t interval of time spent to treat a patient where successive treatment times are identically and independently distributed for a particular physician. Then the mean treatment time representing a continuous density function is

$$\int_0^\infty tG(t)dt = \frac{1}{\psi}. \tag{13}$$

The symbol ψ is the treatment rate per time. In most cases, the service time in queue theory is assumed to be exponentially distributed as earlier stated in model's assumption. Hence, Equation (13) can be stated as:

$$G(t) = \psi\ell^{-\psi t}, \quad t \geq 0. \tag{14}$$

Applying the assumption stated in Section 2, suppose a patient treated at time t was evaluated at time $t + l$, the probability of not completing the treatment in the interval length l can be stated as

$$P[\text{service not completed in interval length of time } l] = \ell^{-\psi l}. \tag{15}$$

Therefore, for a small positive number $l > 0$, the probability that a physician did not complete a treatment in an interval length of time l can be written as

$$P\begin{bmatrix}\text{treatment not completed}\\ \text{in an interval length of time } l\end{bmatrix} \cong 1 - \psi l. \tag{16}$$

While the probability that treatment is completed in an interval length of time l is

$$P\begin{bmatrix}\text{treatment is completed in an}\\ \text{interval length of time } l\end{bmatrix} = \psi l. \tag{17}$$

Suppose there is only one physician with an exponential distribution time density in equation (15) and let the probability that n patients are in the healthcare centre be $P_n(T)$ at time T. Similar to other preceding section, we are to evaluate approximate probabilities and ignore relatively small quantities. Hence, we assume only one patient was treated during a small interval of time $l > 0$. In the case of n patients in the healthcare centre at time $T + l$, we consider two possibilities. Either there were n patients and none of them was treated or there were $n + 1$ patients in the system and one was treated in l small interval of time length.

Consequently, we have

$$P_n(T + l) = (1 - \psi l)P_n(T) + (\psi l)P_{n+1}(T). \tag{18}$$

The first term in RHS of equation (18) represents the probability approximation that no patients was treated in the time interval length l with n patients in the healthcare centre at time T while in the second term, there were $n + 1$ patients and one was treated. Equation (18) can be rearranged to produce Equation (19) as follows:

$$\frac{P_n(T + l) - P_n(T)}{l} = -\psi P_n(T) + \psi P_{n+1}(T). \tag{19}$$

By differentiating Equation (19) and letting $l \to 0$, we have

$$\frac{dP_n}{dT} = -\psi P_n(T) + \psi P_{n+1}(T), \qquad \text{for } 1 \leq n < M. \tag{20}$$

Equation (20) holds exactly because all the small relatively terms ignored in the previous steps vanish as l tends to zero ($l \to 0$). By a similar procedure, we can state that:

$$\frac{dP_m}{dT} = -\psi P_m(T), \quad \text{for } n = M. \tag{21}$$

The differential Equations in (20) and (21) have a unique solution which can be stated as:

$$P_n(T) = \frac{(\psi T)^{-n} \ell^{-\psi T}}{(m-n)!}, \text{ for } n \geq 1. \tag{22}$$

For $n = 0$, Equation (22) becomes

$$P_0(T) = 1 - \sum_{n=1}^{M} P_n\,(T). \tag{23}$$

The distribution in Equations (22) and (23) are referred as truncated Poisson distributions. If the last patient to be treated is M[th] patient, then the sum of time x that the patient spends in the queue system, plus the treatment time will have a gamma density distribution denoted by the sum of variables will have exponential distribution, as stated in Ogumeyo and Emunefe (2022). Thus

$$H(x) = \frac{\psi(\psi x)^{M-1} \ell^{\psi x}}{(M-1)!}, \qquad \text{for } x \geq 0. \tag{24}$$

Hence, the mean and the variance of equation (24) can be stated as

$$E(x) = \frac{M}{\psi} \tag{25}$$

and

$$\text{Var}(x) = \frac{M}{x^2}. \tag{26}$$

Equation (24), (25) and (26) denote gamma distribution for a small interval time $l > 0$, we can apply binomial probability theorem to obtain a small time interval l since the physicians work independently. Thus, the probability that none of the patients departed from the healthcare centre can be mathematically stated as:

$$\text{P[none of the } n \text{ pateints departed]} = (1 - n\psi l)^n = 1 - n\psi h. \tag{27}$$

The probability that one of the patients departed the healthcare centre using equation (27) can be stated thus

$$P[\text{one of the patients departed}] \cong n\psi l. \tag{28}$$

Equation (28) is justified by the fact that when $l > 0$ is very small, we restrict ourselves to only two chances: either there was no departure or one departure. This is because the possibilities of more patients departing the queue system have intangible probabilities. Another reason for this is that, when the healthcare centre has n patients in its queue system at time $T + l$, either there were n patients and none departed or there were $n + 1$ patients and one departed are the only changes that can occur at time T, (see Ogumeyo and Emunefe, 2022).

The above explanation can be expressed mathematically as:

$$P_n(T + l) = (1 - n\psi l) P_n(T) + (n + 1)\psi l P_{n+1}(T). \tag{29}$$

We arrange Equation (29) by taking the term P_n(T) to the LHS. Divide both sides the LHS by l and allowing l tends to zero, we have

$$\frac{dP_n}{dT} = -n\psi P_n(T) + (n + 1)P_{n+1}(T), \quad 0 < n < M. \tag{30}$$

By following the same procedure, we have

$$\frac{dP_M}{dT} = -M\psi P_M(T), \text{ when } n = M. \tag{31}$$

By using Binomial distribution probability, the complete solution of Equations (30) and (31) can be stated thus:

$$P_n(T) = \binom{M}{n}(\ell^{-\psi T})^n(1-\ell^{-\psi T})^{M-n}, \quad \text{for } 0 < n \leq M. \tag{32}$$

The mean of the distribution of the queue system using equations (30)-(32) can be stated as

$$E\left[\frac{n}{T}\right] = M\ell^{-\psi T}. \tag{33a}$$

While the variance is

$$\text{Variance}\left[\frac{n}{T}\right] = M\ell^{-\psi T}(1-\ell^{-\psi T}). \tag{33b}$$

Model 2: Mathematical Formulation

Model description: In model 2, we use Kolmogorov differential equations to derive the transient solution, the statistical equilibrium and the stationary probabilities of the single-server model described in Model 1. Specifically, we assume that

$$\begin{aligned}\text{exponential density of interarrival time } &= \lambda e^{-\lambda t}\\ \text{exponential density of service time } &= \mu e^{-\mu t}\end{aligned} \tag{34}$$

$\lambda =$ patients' arrival rate per time,

where E[Interval of Busy Period] $= E$[Time patients spent in the queue system]

$\mu =$ treatment rate of patients per time, $\rho = \lambda/\mu$ traffic intensity.

The number of patients, *n* in the queue system at any given time consists of patients waiting in the queue including those currently in service. Suppose we denote $t = 0$ as the time the queue system started, then $P_n(T)$ can be defined as

$$P_n(T) \equiv \text{ the chances that } n \text{ patients are currently queuing in system at time } T. \tag{35}$$

The value of $P_n(T)$ is determined by the number of patients in the queue system at time $t = 0$. Suppose $k > 0$ is a *very* small length of time where n greater than 0 is the number of patients in the queue system at Time $T + k$, then similar to the procedure in Model 1, we consider only the chances that there were either $n-1, n$ or $n+1$ patients at time T while any other chance is relatively negligible. Hence, for n greater than zero, we have

$$\begin{aligned}\mathrm{P_n}(T+k) \equiv\ &(\lambda k)(1-\mu k)P_{n-1}(T) + (1-\lambda k)(1-\mu k)P_n(T) + (\lambda k)(\mu k)P_n(T)\\ &+ (\lambda k)(\mu k)P_n(T) + (1-\lambda k)(\mu k)P_{n+1}(T) \qquad (\text{for small } k)\end{aligned} \tag{36}$$

The first term on the right hand side represents the arrival of one patient and no patient departed the system when there are $n-1$ patients in the queue at time T. The second and third terms in right hand side of (36) represent the events: no patient arrives and no patient departs, and of one arrival and a departure when there are *n* patients in the queue at time T respectively. The last term denotes the event of no arrival and one departure when there are $n+1$ patients in the queue system at time T. The expression " $\equiv$ " in (36) is a sign of approximation which can become "=" by adding probability terms with coefficient k^i, where $i \geq 2$. The probability $P_n(T)$ that the system has exactly n patients either waiting for treatment or in treatment process at time t satisfies the Kolmogorov differential equation in (36). If we shift the term $P_n(T)$ to the LHS of Equation (36), divide through by k, and letting $k \to 0$, we have

$$\frac{dP_n}{dT} = \lambda P_{n-1}(T) - (\lambda + \mu)P_n(T) + \mu P_{(n+1)}(T) \quad \text{for } n > 0. \tag{37}$$

Since the terms we ignored in Equation (36) become zero as $k \to 0$, the expression in equation (37) is exact and not an approximation. Hence, for $n = 0$, we have

$$\frac{dP_0}{dT} = -\lambda P_0(T) + \mu P_1(T) \quad \text{for } n = 0. \tag{38}$$

Before we can solve linear differential equations in (37) and (38) for each value of $P_n(T)$, we need to know how many patients (n) are in the queue system at time $t = 0$. This leads us to arrive at a transient solution since its value solely depends on the value of T. We can also examine the values of $P_n(T)$ as T tends to infinity (∞). If $P_n(T)$ approaches a limiting value P_n, and $E[n]$ is finite for this limiting distribution then, the queue system is said to have attained its statistical equilibrium. The P_n values obtained shows that if the number of patients in the system at any time t is specified with respect to the probability distribution P_n, then for any $k > 0$, P_n is also the probability that n patients are in the queue system at time $t + k$. The P_n value can also denote the value of the limit of an arbitrarily long period of time in which the queue consists of n patients provided:

$$\rho \equiv \frac{\lambda}{\mu} < 1. \tag{39}$$

The stationary probabilities P_n are common features in single-server $(M/M/1)$ and the symbol ρ in (39) represents the traffic intensity. By using the rule which states that each dP_n/dT must equal 0, for P_n solution to be independent of T, the equilibrium solution, $P_n(T) \equiv P_n$ can be determined for all T. (See Ogumeyo and Emunefe, 2022). Hence, in order to evaluate the P_n values, we need to set the derivatives in (37) and (38) equal to 0, to get:

$$0 = \lambda P_{n-1} - (\lambda + \mu)P_n + \mu P_{n+1} \quad \text{for } n = 1, 2, 3, \ldots \tag{40}$$

$$0 = -\lambda P_0 + \mu P_1 \quad \text{for } n = 0. \tag{41}$$

Starting with (41), the system of equations in (40) and (41) can be solved recursively. Thus,

$$P_1 = P_0\left(\frac{\lambda}{\mu}\right) = P_0\rho\,. \tag{42}$$

If we proceed to (40) for $n = 1, 2, \ldots$, we have

$$P_n = P_0\rho^n. \tag{43}$$

We can easily verify that P_n in (43) does satisfy (40). Given Equation (39),

$$\sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \rho^n = \frac{P_0}{1-\rho} = 1. \tag{44}$$

Equation (44) implies that $P_0 = 1 - \rho$, hence

$$P_n = (1-\rho)\rho^n \quad \text{for } n = 0,1,2,\ldots. \tag{45}$$

Equation (45) denotes a geometric distribution, hence

$$E\begin{bmatrix}\text{number of}\\ \text{patients}\\ \text{in system}\end{bmatrix} = E[n] = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}, \qquad \text{var}[n] = \frac{\rho}{(1-\rho)^2}. \tag{46}$$

Equation (46) holds for $P[n \geq N] = \rho^N$.

We observe that the probability distribution in (45) rely only on the value of $\lambda/\mu = \rho$, which is the traffic intensity ratio. Hence ρ is the measure of effectiveness or utilization factor of the queue system since $\rho(=1-P_0)$ is the time interval in which the physician is busy.

Derivation of Mean of Queue Length

In order to derive the mean of the queue length, we first all recall that

$$\text{Queue length} = \begin{cases}\text{number of patints in the queue system} & \text{if } n = 0\\ \text{number of patients in the queue system } -1 & \text{if } n > 0.\end{cases} \tag{47}$$

Hence,

$$E[\text{Queue length}] = 0.P_0 + \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=0}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n \tag{48}$$

$$= E[n] - (1-P_0) = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-2)} .$$

Derivation of the Length of Idle and Busy Periods

Next step is to study the intervals of time in which the physician is idle which normally starts when one treatment terminates and end when a new patient arrives. The inter-arrival times of patients and the interval of the idle period of the physician have the same distribution. That is, exponentially distributed with mean $= 1/\lambda$. Suppose the time interval T is long enough that its expected values can be fully utilize, then the physician will be idle for $TP_0 = T(1-\rho)$ time units, and $[T(1-\rho)/(1/\lambda) = \lambda T(1-\rho)]$ will be the number of different idle periods in time T. This implies that $\lambda T(1-\rho)$ is also the number of different busy periods in time T since idle and busy periods alternate, and ρT is the sum ofthe time interval in which the physician is busy.

Hence,

$$E[\text{interval of busy period}] = \frac{\rho T}{\lambda T(1-\rho)} = \frac{1}{\mu-\lambda} \tag{49}$$

and

$$\text{E}\begin{bmatrix}\text{number of patients}\\ \text{treated per busy period}\end{bmatrix} = \mu \text{E}\begin{bmatrix}\text{interval of}\\ \text{busy period}\end{bmatrix} = \frac{1}{1-\rho}. \tag{50}$$

Equations (49) and (50) hold for any service time distribution, (see Ogumeyo and Emunefe, 2022).

Next, we consider the density probability for the time a patient waits in the queue system, which consists of the length of time a patient stays in the queue and the time he spend in treatment process. Assuming the queue system is in a state of statistical equilibrium, such that when a new patient arrives, there is a probability P_n given by Equation (45) of finding n patients in the system ahead of him/her and the first come, first served queue discipline is observed, then the total time the patient stays in the queue system will be the sum of $n+1$ identically and independently distributed exponential random variables. Hence, it will have a gamma density

$$h(\omega)\frac{\mu(\mu y)^n e^{-\mu y}}{n!} \quad \text{for } y \geq 0. \tag{51}$$

Equation (51) is similar to what we have in Equation (48), hence the time interval that a patient who comes at an arbitrary time stays in the queue system is given by

$$h(\omega) = \sum_{n=0}^{\infty}(1-\rho)\rho^n\left[\frac{\mu(\mu\omega)^n e^{\mu\omega}}{n!}\right] \tag{52}$$

$$= \mu(1-\rho)e^{-\mu(1-\rho)w} \quad \text{(which is exponentially distributed).}$$

While mean time in system is

$$E[\text{time in the queue system}] = E[\omega] = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu-\lambda}. \tag{53}$$

Recall that $E[\text{time in line}] = E[\text{time in system}] - E[\text{service time}]$, hence

$$E[\text{time in line}] \quad = \frac{1}{\mu}\left(\frac{\rho}{1-\rho}\right) = \frac{\lambda}{\mu(\mu-\lambda)}. \tag{54}$$

From the above analysis of Equations (49)-(54), it implies that for fixed value of ρ, the mean time a patient stays in the queue system and in the queue inversely vary with the treatment rate μ. Difference equations of the steady-state in equations (40) and (41) still hold for $n = 0, 1, \ldots, M-1$, but the equation for $n = M$ is

$$0 = \lambda P_{M-1} - \mu P_M. \tag{55}$$

While the solution corresponding to the values of $n = 0, 1, \ldots, M$ is

$$P_n = \begin{cases} \left(\frac{1-\rho}{1-\rho^{M+1}}\right)\rho^n & \text{for } \lambda \neq \mu \\ \frac{1}{M+1} & \text{for } \lambda = \mu. \end{cases} \tag{56}$$

This also implies that, when $\lambda < \mu$ and $M \to \infty$, P_n in Equation (56) agrees with Equation (35). By applying simple algebraic manipulations, we can be prove that for $\lambda \neq \mu$

$$E\begin{bmatrix}\text{number of} \\ \text{patients} \\ \text{in system}\end{bmatrix} = E[n] = \frac{\rho}{(1-\rho)}\left[\frac{1-(M+1)\rho^M + M\rho^{M+1}}{1-\rho^{M+1}}\right] \tag{57}$$

$$= \frac{\rho}{1-\rho} - \frac{(M+1)\rho^{M+1}}{1-\rho^{M+1}} \quad \text{for } \lambda \neq \mu.$$

We notice that when $\lambda < \mu$, the mean number of patients in the queue system is less than that of infinite queue length in Equation (46). By following the same procedure, we can prove that for $\lambda = \mu$,

$$E\begin{bmatrix}\text{number of} \\ \text{patients} \\ \text{in system}\end{bmatrix} = E[n] = \frac{M}{2} \quad \text{for } \lambda = \mu. \tag{58}$$

Special case of patients who arrive when the system is already full

Determining the length of time a patient stays in a queue system is a sensitive aspect of queuing theory. Patients who arrive at the health facility when there is no vacant space in the system do not enter the system due to lack of space. This implies that the patient waste no time in the system. Hence, the mean time spent in the system is defined as either all customers who come regardless of whether they enter, or only to those patients who are allowed to enter. In this case the latter is adopted, since in most cases attention is focused on only those who actually enter the system. Hence, for a patient who arrive at an arbitrary time to join the system, with first come, first served queue discipline being observed, we can state that

$$E\begin{bmatrix}\text{time in}\\ \text{system}\end{bmatrix} \equiv E[w] = \frac{\rho}{\mu(1-\rho)}\left[\frac{1-\mu\rho^{M-1}+(M-1)\rho^{M}}{1-\rho^{M}}\right]+\frac{1}{\mu} \tag{59}$$

$$= \frac{1}{\mu(1-\rho)} - \frac{M\rho^{M}}{\mu(1-\rho^{M})} \quad \text{for } \lambda \neq \mu,$$

and

$$E\begin{bmatrix}\text{time in}\\ \text{system}\end{bmatrix} \equiv E[w] = \frac{1}{\mu}\cdot\frac{M+1}{2} \quad \text{for } \lambda = \mu. \tag{60}$$

Numerical Illustration

Different operating characteristics such as expected number of patients in line, expected line length and time sent in line in a single-server queuing (M/M/1) system for different values of traffic intensity ρ and service rate μ are displayed in Table 1 as follows:

Traffic Intensity ρ	Probability of Server Idle $=1-\rho$	Expected number in system $=\frac{\rho}{1-\rho}$	Expected Line Length $=\frac{\rho^2}{1-\rho}$	$\mu = 10$			$\mu = 20$		
				λ	Time in systems	Time in line	λ	Time in system	Time in Line
0.1	0.9	0.11	0.01	1	0.11	0.01	2	0.06	0.01
0.3	0.7	0.43	0.13	3	0.14	0.04	6	0.07	0.02
0.5	0.5	1.00	0.50	5	0.20	0.10	10	0.10	0.05
0.7	0.3	2.33	1.63	7	0.33	0.23	14	0.17	0.12
0.8	0.2	4.00	3.20	8	0.50	0.40	16	0.25	0.20
0.9	0.1	9.00	8.10	9	1.00	0.90	18	0.50	0.45
0.95	0.05	19.00	18.05	9.5	2.00	1.90	19	1.00	0.95
0.99	0.01	99.00	98.01	9.9	10.00	9.90	19.8	5.00	4.95
0.999	0.001	999.00	998.00	9.99	100.00	99.90	19.98	50.00	49.95

$\lambda =$ rate of patients' arrival per time, $E[\text{length of Busy Period}] = E[\text{Time in System}]$

$\mu =$ rate of treatment per time, and $\rho = \lambda/\mu$

5. Results and Analysis

In model 1, the inter-arrival time distribution of patients in a healthcare centre is being described in Equations (1) to (12). The model is based on the assumption that the time intervals between patients' arrivals and the treatment time have identical and independent distribution, denoting a density function that is continuous and the patients' arrival is a renewal process. This enables us to evaluate approximate probabilities of arrival time and ignoring very small-time quantities. Hence, in a very small-time interval, $l > 0$, it can be stated that at most only a single patient can be treated (i.e. one departure). When there are n patients at time $T + l$ in the system, we consider only two probabilities: either there were n patients in the queue system and none have left or there were $n + 1$ patients and one has left the facility within the very small length of time $l > 0$. This gives the results in Equations (13) to (18). We observe that Equation (19) has exact probability, and not approximate. This is because all the terms with very small probabilities that were overlooked in Equation (18) by letting $l \to 0$ disappear. Equations (25) and (26) show the mean and variance of the total time a patient spent in the queue including the treatment time using Gamma distribution, respectively. Moreover, Equations (28) to (31) are justified by the fact that whenever an interval of time $l > 0$ is very small, we are restricted to only two possible chances: either no patient left the system or one patient left the system since the probability that more patients will leave the system is very small and negligible. Equations (33) contain the mean and variance of the Binomial distribution of the service or treatment rate of patients.

Table 1 shows that as the value of ρ increases, the mean number of patients, the queue length, the time spent in the system, and the time in the queue [using Equations (46), (48), (53) and (54)] stated above, all increase rapidly. For sufficiently large $\rho < 1$, these quantities can be made arbitrarily large. Although it will take the system a long time to reach steady-state equilibrium. For a specified treatment rate μ, when ρ is small, most of the mean times a patient spends in the system is caused by the average service time $1/\mu$; but as the arrival rate λ increases (intensity ρ increases) most of the expected time spent in the system is caused by waiting time in the queue.

6. Conclusion

In this paper, we presented two queuing models aimed at solving problem of long queues in hospitals. In model 1, we apply the queue discipline approach of first-come first-served to derived the distribution probability that govern the numbers of patients' arrival and departure in any given interval, where $\emptyset$ denotes the average patients' arrival rate while ψ is the average rate of treatment of patients or service rate. We observed that the distribution probability of inter-arrival time and treatment time of patients describes a continuous density function similar to that of a renewal process if the expected inter-arrival and expected treatment times of a patient are known. In model 2, we used the method of Komolgorov linear differential equations (37) and (38) for each $P_n(T)$, to derive the transient solution. The values of $P_n(T)$ is also examined as T tends to infinity (∞) and we observed that the queue system can reach its statistical equilibrium state, if $P_n(T)$ tends to a limit, say P_n, and $E[n]$ is finite for the limit P_n. The resulting P_n values (i.e. stationary probabilities) imply that if the number of patients in the system at any time t is given with respect to the probability distribution P_n, then for any $k > 0$, P_n is also the probability that *n* patients are in the system at time $t + k$. We also observed that the equilibrium solution $P_n(T) \equiv P_n$, can be obtained for all T, by equating each $dP_n/dT = 0$, if the solution P_n is truly independent of T. (See Ogumeyo and Emunefe, 2022).

Table 1 shows that as the value of ρ increases, the mean number of patients, the queue length, the time spent in the system, and the time in the queue [using Equations (46), (48), (53) and (54)] stated above, all increase rapidly. For sufficiently large $\rho < 1$, these quantities can be made arbitrarily large. Although it will take the system a long time to reach steady-state equilibrium. For a specified treatment rate μ, when ρ is small, most of the mean times a patient spends in the system is caused by the average service time $1/\mu$; but as the arrival rate λ increases (intensity ρ increases) most of the expected time spent in the system is caused by waiting time in the queue. Our proposed models have advantage over the existing ones in that they contain mathematical formulas that could be used to determine the optimum values of transient solution, statistical equilibrium and stationary probabilities of a queue system. These values could guide hospital management personnel to make policies which shorten queues due to prompt services and enhance the sustainability of patients' interest and patronage. This in effect, leads to high revenue generation and profit maximization. Another advantage of these two models is the prevention of stampede which could lead to loss of lives such as was experienced in Ibadan, Abuja and Okija (all in Nigeria) during the distribution of palliatives in December, 2024.

References

[1] Abou-El-Ata, M. O. M., & Hariri, A. M. A. (1992). The M/M/c/N queue with balking and reneging. *Computers & Operations Research, 19*, 713-716. https://doi.org/10.1016/0305-0548(92)90010-3

[2] Afrane, S., & Appah, A. (2014). Queuing theory and management of waiting-time in hospitals: The case of Anglo Gold Ashanti Hospital in Ghana. *International Journal of Academic Research in Business and Social Sciences, 4*(2), 33-44. https://doi.org/10.6007/IJARBSS/v4-i2/590

[3] Ba, A. K., & Farnaza (2017). An assessment of patients' waiting and consultation time in a primary healthcare clinic. *International Journal of Academic Research, 12*(1), 14-21.

[4] Belson, G. V. (1988). Waiting times for scheduled patients in the presence of emergency requests. *Journal of Operational Management, 2*(3).

[5] Boudali, O., & Economou, A. (2007). Equilibrium customer strategies in a single-server Markovian queue with setup times. *Queueing Systems, 56*(3-4), 213-228. https://doi.org/10.1007/s11134-007-9036-7

[6] Cabolat, P. G. (2020). Bounded rationality in clearing service systems. *European Journal of Operational Research, 282*(2), 614-626. https://doi.org/10.1016/j.ejor. 2019.10.013

[7] Chai, X., Liu, L., Chang, B., Jiang, T., & Wang, Z. (2019). On a batch matching system with impatient servers and boundedly rational customers. *Applied Mathematics and Computation, 354*, 308-328. https://doi.org/10.1016/j.amc.2019.02.004

[8] Economou, A., & Kanta, S. (2008). Optimal balking strategies and pricing for the single-server Markovian queue with compartmented waiting space. *Queueing Systems, 59*, 237-269. https://doi.org/10.1007/s11134-008-9083-8

[9] Falin, G. I., & Templeton, J. G. C. (1997). *Retrial queues.* Chapman & Hall. https://doi.org/10.1007/978-1-4899-2977-8

[10] Green, L. (2006). Queueing analysis in healthcare. In R. W. Hall (Ed.), *Patient flow: Reducing delay in healthcare delivery* (pp. 281-307). Springer. https://doi.org/10.1007/978-0-387-33636-7_10

[11] Hassin, R. (2016). *Rational queuing*. CRC Press. https://doi.org/10.1201/b20014

[12] Haviv, M., & Kerner, Y. (2007). On balking from an empty queue. *Queueing Systems, 55*, 239-249. https://doi.org/10.1007/s11134-007-9020-2

[13] Hogarth, R., & Haviv, M. (2003). *To queue or not to queue: Equilibrium behavior in queuing systems*. Kluwer Academic Publishers.

[14] Huang, T., & Chen, Y. J. (2015). Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management, 24*(5), 778-790. https://doi.org/10.1111/poms.12298

[15] Huang, T., Allon, G., & Bassamboo, A. (2013). Bounded rationality in service systems. *Manufacturing & Service Operations Management, 15*(2), 263-279. https://doi.org/10.1287/msom.1120.0417

[16] Jain, N. K., Kumar, R., & Kutnar Som, B. (2014). An M/M/1/N queuing system with reverse balking. *American Journal of Operations Research, 4*, 17-20.

[17] Kumar, B. K., Parthasarathy, P. R., & Sharafali, M. (1993). Transient solution of an M/M/1 queue with balking. *Queueing Systems, 13*, 441-448. https://doi.org/10.1007/BF01149265

[18] Li, X., Guo, P., & Lian, Z. (2016). Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management, 25*(11), 1885-1901. https://doi.org/10.1111/poms.12583

[19] Li, X., Guo, P., & Lian, Z. (2017). Price and capacity decisions of service systems with boundedly rational customers. *Naval Research Logistics, 64*(6), 437-452. https://doi.org10.1002/nav.21767

[20] Lohlun, K. N., Kotzen, J. A., & Lakier, R. (2015). A prospective study on the impact of waiting times for radiotherapy in cervical cancer. *South African Journal of Obstetrics and Gynecology, 21*(1), 6-9. https://doi.org/10.7196/sajog.985

[21] Mahala, V., Patel, J. D., & Zaveri, B. (2023). A study on using queuing theory to reduce OPD waiting time in hospital operations. *International Journal of Innovative Science and Research Technology, 8*(2), 1409-1413.

[22] Montazer-Haghighi, A., Medhi, J., & Mohanty, S. G. (1986). On a multi-server Markovian queuing system with balking and reneging. *Computers & Operations Research, 13*, 421-425. https://doi.org/10.1016/0305-0548(86)90029-8

[23] Natvig, B. (1975). On a queuing model where potential customers are discouraged by queue length. *Scandinavian Journal of Statistics, 2*, 34-42.

[24] Nithya, R. P., & Haridass, M. (2016). Analysis of a queuing system with two phases of bulk service, closedown and vacation interruption. *International Journal of Applied Engineering Research, 11*, 467-468.

[25] Paul, B. C., Kumar, N., Kumar, A., & Singh, I. B. (2021). Queuing model to optimize patient-waiting time in outpatient department of a public hospital in India. *Journal of Emerging Technologies and Innovative Research, 8*(11), 677-686.

[26] Ogumeyo, S. A., & Emenonye, C. A. (2023). Analytical behavior of customers and their costs' implications in queuing systems. *Journal of Scientometrics and Innovations*, *1*(1), 140-148.

[27] Ogumeyo, S. A., & Emunefe, J. O. (2022). A queuing model to analyze the probability distributions of vehicles' inter-arrival and service times in a traffic intersection. *Journal of the Nigerian Association of Mathematical Physics, 64*, 105-110.

[28] Rajadurai, P. (2018). Sensitivity analysis of an M/G/1 retrial queuing system with disaster under working vacations and working breakdowns. *RAIRO: Operations Research, 52*, 35-54. https://doi.org/10.1051/ro/2017091

[29] Ren, H., & Huang, T. (2018). Modeling customer bounded rationality in operations management: A review and research opportunities. *Computers & Operations Research, 91*, 48-58. https://doi.org/10.1016/j.cor. 2017.11.002

[30] Singh, C. J., Jam, M., & Kumar, B. (2014). Analysis of Mx/G/1 queuing model with balking and vacation. *International Journal of Operational Research, 19*, 154-173. https://doi.org/10.1504/IJOR.2014.058952

[31] Singh, S. N., & Tiwari, S. B. (2013). An application of generalized entropy in queuing theory. *Journal of Applied Sciences and Engineering, 16*, 99-103.

[32] Sudhesh, R., Azhagappan, A., & Dharmaraja, S. (2017). Transient analysis of M/M/1 queue with working vacation, heterogeneous service, and customers' impatience. *RAIRO: Operations Research, 51*, 591-606. https://doi.org/10.1051/ro/2016046

[33] Sztrik, J. (2012). *Basic Queuing Theory*. Faculty of Informatics, University of Debrecen, Hungary.

[34] Wang, J., & Zhang, Z. G. (2013). Strategic joining in M/M/1 queue with risk-sensitive customers. *Journal of the Operational Research Society, 69*(8), 1197-1214. https://doi.org/10.1080/01605682.2017.1390526

[35] Wang, J., Zhang, Y., & Zhang, Z. G. (2021). Strategic joining in an M/M/k queue with asynchronous and synchronous multiple vacations. *Journal of the Operational Research Society, 72*(1), 161-179. https://doi.org/10.1080/01605682.2019.1644978

[36] Worthington, D. J. (1987). Queuing models for hospital waiting lists. *Journal of the Operational Research Society, 38*, 413-422. https://doi.org/10.1057/jors.1987.69

[37] Zhang, Y., Wang, J., & Wang, F. (2016). Equilibrium pricing strategies in retrial queuing systems with complementary services. *Applied Mathematical Modeling, 40*(11-12), 5775-5792. https://doi.org/10.1016/j.apm.2016.01.029