How to Cite:
Mittal, H., & Sharma, N. (2022). A simulation-based approach for minimizing waiting time in AIIMS, Delhi using Queuing model. *International Journal of Health Sciences*, *6*(S5), 7037–7054.
https://doi.org/10.53730/ijhs.v6nS5.10227

A simulation-based approach for minimizing waiting time in AIIMS, Delhi using Queuing model

Himanshu Mittal
School of Engineering and Sciences, GD Goenka University, Gurugram, Haryana, India
Email: himanshu1000mittal@gmail.com

Naresh Sharma
School of Engineering and Sciences, GD Goenka University, Gurugram, Haryana, India
Corresponding author email: naresh.sharma2006@gmail.com

***Abstract*---**The government hospitals in India influenced by multiple factors causing longer waiting time of patients in comparison to private hospitals, which worsen the threat to healthcare facilities. The optimization of available resources considering the arrival rate of patients and the availability of facilities for the minimization of queuing is the utmost requirement. The current study has been carried out within the outpatient's department to minimize queue length of one of the largest and busiest hospital, AIIMS in Delhi, represent a struggling health care delivery system with high waiting times of patients. The primary queuing data was collected for total samples of 1200 patients during the four-week study period (1st July to 30th July 2020), (Monday-Friday) and working hours of general OPD (08:30 am to 01:00 pm). The detailed queuing secondary data was collected from AIIMS for three years (1st January 2015 to 31st July 2017). Data has been analyzed by queuing models, M/M/1: Poisson-exponential, single server model-infinite population and up to M/M/8: Poisson-exponential, multiple server model-infinite populations.

***Keywords*---**Queuing model, Optimization of Queues, AIIMS, single server model and multiple server model.

1 Introduction

A significant criterion for efficiency measurement within the service industry is the waiting time (Tadj, 1996; Sharma et. al. 2010, 2011). The issue of waiting times of patients and queue length assessment has been widely researched in

healthcare (Barlow, 2002; Obulor and Eke, 2016; Ibrahim, 2017; Mittal and Sharma, 2020b); however, maximum studies have been conducted in the developed world, and very few studies have analyzed the queuing systems in India. The queuing problem is immensely crucial in the healthcare services, due to the involvement of high morbidity and mortality rate of patients.

When contemplating changes in infrastructure, the healthcare manager balances the costs of offering a specific level of service against the future costs of waiting for treatment. The two core costs referred to here are the costs associated with waiting time for patients who may be unable to wait for the service and choose to go to some other hospitals, costs associated with the delay in care and value of the amount of time spent for patients and the reduction in patients satisfaction. Service costs include staff salaries and employee benefits or servers waiting to be served on other servers. By using the cost estimate, decision-makers can determine the optimal server number, reducing the total costs, including the cost of the service and the cost of waiting. The waiting costs for each person depending on what the person earns each hour. A.K Erlang first analyzed the queuing theory in 1913. The theory has been used in many fields of industry and public service since the Second World War (Kalashnikov, 1990; Lee et al., 1997). Queuing literature notes that waiting in line or queue creates economic costs to individuals and organizations. Healthcare, airlines, banks, manufacturing firms and other businesses are seeking to reduce the overall waiting costs and customer support costs. The researcher worked on the modelling of impact on the use, the waiting time, and the likelihood of refraining patient bed assignment policies. Several studies are carried out to investigate customer satisfaction using Queuing theory. Excessive waiting and service times worsen the patient demand, increases the cost of health care, constitute a barrier to effective treatment, and causes dissatisfaction among the patients (Thomas, 2013) where they might leave the system without receiving the service. Due to increasing competition, private players in healthcare services strive to provide fast and efficient health services to attract more patients. All India institute of medical science (AIIMS), Delhi also teamed up with Indian Statistical Institute (ISI) to intelligently and accurately predict the patient turn up the rate to every department by scientifically analyzing the historical turn-up data from e-hospital (AIIMS, 2020). In the queuing system, by considering a stream of arrivals, arriving at a service node, within which they are processed, and then exiting from the system shows in Figure 1. The node can only take a certain number of arrivals and process them at the same time.

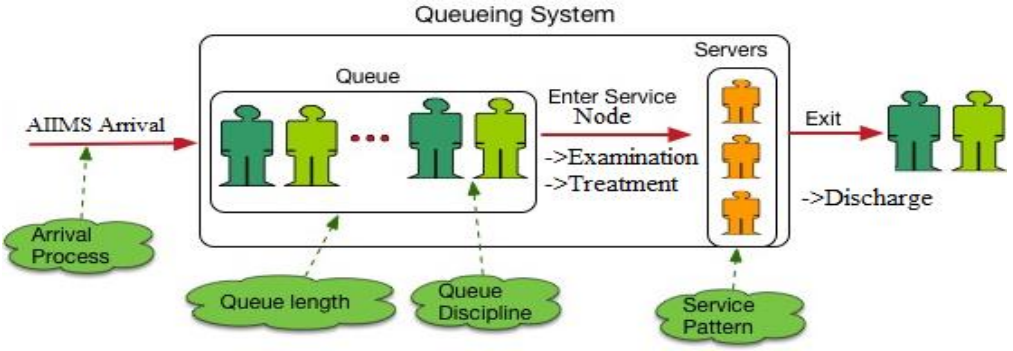

Figure 1: Schematic representation of Queuing system in AIIMS, Delhi

2 Literature Review

The analysis of waiting time in healthcare systems can be approached mathematically using queuing theory (Rotich, 2016). Queuing theory is a classical approach that has been employed to gain insight into hospital processes since the 1950s (Lee et al., 1997; Titovtsev, 2016). An ineffective and inefficient appointment system is one of the main issues for the long waiting time of patients in an outpatients department. Therefore, various studies have been conducted which aim at improving the appointment scheduling in an outpatients' department. Nuyens (2004) investigated the average as well as the maximum length of the queue in an outpatients' department, and assessed their relationship with other queuing factors such as average wait time, an average number of patients served and maximum idle and busy time for doctors. The author considered the queue length for one particular examination room and evaluated its relationship with the arrival rate of patients (Mittal and Sharma, 2022b). They concluded that although queue length increases as arrival rate escalate, it is not overwhelming. Akilan (2011) proposed changes to the existing appointment schedule by investigating the effect of overtime, no-show rate and overbooking on congestion and increased length of stay. Lefeber et al. (2011) developed several appointment systems by incorporating different combinations of various appointment rules and patient sequences. These appointment systems were tested for punctual patients as well as for no-shows. Kirpichnikov and Titovtsev (2016) provide particularly focused on identifying multiple elements which led to long queues and testing different appointment systems to identify an optimal system.

Bako et al. (2017) developed a model of determining optimal rules for an outpatient appointment system. These factors included the uneven distribution of appointment slots, late start of sessions, unused sessions when no patients are seen and irregular calling sequence of appointment patients for consultation. The length of the queue has explicitly been included as a significant queuing variable in some studies. Setiawan and Nugraheny (2019) specifically considered routine and urgent patients in order to analyze and improve the current outpatient appointment scheduling, by assessing the queue length, wait time, and physician's idle time and overtime. Yaduvanshi et al. (2019) stated that if this property does not obtain, or cannot be approximated within a fair degree of accurateness, then queuing theory is unable to provide a systematic solution to the waiting-time problem unless priorities do not play an essential role in the service mechanism. Joseph (2020) included length of the queue as a measure of congestion in an emergency department, particularly considering those patients who left without being seen due to excessive waits. They evaluated the queue length and wait times for two types of queues, including the queue for treatment and queue for bed in the ward.

2.1 Motivation and objective of the study

In this current study, data has been collected from the outpatients' department of one of the largest and busiest government hospitals in India, i.e. AIIMS, Delhi, a representative of struggling health care delivery systems with high waiting times of patients. Appointment system and overcrowding was another major issue faced

by AIIMS OPD. The large number of patients arrive daily along with the care takers/relatives make the place over crowded (AIIMS, 2020) and affects the functioning of the system. The minimization of queues in AIIMS, Delhi health facilities affected by several factors causing higher waiting time of patients, and several studies reflect that waiting time in AIIMS, Delhi is relatively longer in comparison to private hospitals. The various waiting times are also experienced in different service categories in health facilities, with studies indicating that an emergency section tends to have short waiting times. Analytic solution of the model relies on the Poisson nature of the arrival process. If this property does not obtain, or cannot be approximated within a fair degree of accurateness, then queuing theory is unable to provide an analytic solution to the waiting-time problem unless priorities do not play an essential role in the service mechanism (Mittal and Sharma, 2020a).

Queuing is an essential criterion of efficiency assessment which represents the operational capability of large busy government hospitals, with excessive queuing causing inconvenience for the patients. However, besides continuously striving to achieve this goal, there is a dire need to construct a framework which could guide the management of the busy hospital to minimize the excessive waiting by patients, in the absence of prior appointments, given the current meagre conditions. The study carried out to minimize the queue length of AIIMS, Delhi, with the optimization of resources.

2.2 The rationale of the study

The problem of prescribing waiting times of patients was studied where the patient's long waiting time is described by a state space, X. In practice, X may consist of a set of vectors whose components correspond to various measures of prolonged wait times (high patient load, time of arrival, age, few doctors, day of arrival, treatment-related diagnosis, providers, record clerks and clinicians, etc.).
In this case, to trace the problem, we need to assume an ordering of these states (vectors). In doing so, we lose no generality in also assuming that states can be represented by natural (scalar) numbers. For simplicity, we model X: = {1, 2, …, n} for some positive integer n ≥ 2. Here we take state 1 to be the best state of health and state n to be the worst. The practitioner (e.g. a physician) has a finite number, k, of possible treatments to prescribe to the patient during each appointment. Let A: = {1, …, k} denote the set of treatments. Associated with each treatment a ∈ A is a waiting time w(a), and two rates λ(a) and μ(a) where(λ): Mean arrival rate, (μ): Service Rate. The treatment queue function w: A ↦ R+ is assumed to be independent of the patient's state of the queue, and represents the entire queue of a particular treatment incurred from the time of prescription until the next appointment. The rate functions λ: A ↦ R+ and μ: A ↦ R+ are also assumed to be independent of the patient's waiting time and represent treatment response.

More specifically, $\lambda(a)$ can be thought of as the rate at which the patient is waiting time worsen to the next highest state when undergoing treatment 'a'. Similarly, $\mu(a)$ is interpreted as the rate at which the patient's service rate improves to the next lowest state while using treatment 'a'. We assume

that the treatments are ordered by (increasing) Service Rate and effectiveness. That is, the author assumes

$$w(1) \leq w(2) \leq \ldots \leq w(k)$$

$$\lambda(1) \geq \lambda(2) \geq \ldots \geq \lambda(k)$$

$$\mu(1) \leq \mu(2) \leq \ldots \leq \mu(k) \qquad (1)$$

Given the dynamics of this queuing model, there is an immediate tradeoff between the patient's overall state of service rate (μ) and mean arrival rate (λ).

3 Method and Methodology

The current study has addressed a significant issue by highlighting the suffering of patients in terms of long wait times at a typical overcrowded government hospital (AIIMS, Delhi) in India. Queuing theory have real applications to reduce wait-time of a queue, emergency department service system, etc. (Thomas, 2013; Mittal and Sharma, 2022a). In Kendall's notation A/S/c(Arrival/Service/Channel), inter-arrival times that are exponentially distributed are defined by an M (referring to the Markovian property of the distribution) (Kalashnikov, 1990). Thus, we write M/M/c, which in short yields the distribution between first the arrivals, then the finished services, and last the number of parallel servers (Zhang et al., 2014). In this paper, a single server model is used the notation M /M /1 to characterize the priority queuing system in the queuing model.

Table1: Model formulation criteria

S.No.	Model Factors	Parameters
1	Set up	Hospital
2	Input	Arrival of Patients
3	Output	Relieving of Patients
4	Mechanism	Services
5	Queue Discipline	First-Come-First-Serve (FCFS)

3.1 Markov Processes

A random current state on how it arrived in the current state is said to be a Markovian process. The mean inter-arrival times and mean arrival rates may be deterministic or probabilistic. Either of these measures suffices in describing the arrival pattern.

Arrival: Mean arrival rate and is denoted by A(t). Where $A(t)$=Prob [the time between arrivals ≤t] (Chang, 2019).

Service: Service patterns may be deterministic or stochastic. In case these are stochastic, the service time distribution, like the inter-arrival distribution, also follows some type of probability distribution denoted by $B(t)$, that is $B(t)$=Pr*ob* [*service time* ≤ *t*] (Chang, 2019).

Markovian Property: One of the unique features of the exponential distribution is the 'memoryless' or 'forgetfulness' or 'Markovian' or property**.**

$$A(t)=\lambda e^{-\lambda t} \quad , t \geq 0 \qquad (2)$$

The conditional probability of arrival during the time interval (t, t + dt) given that arrival has not taken place in $[0,t_0]$ is defined as

$$\text{Prob}\left\{t<T<t+\frac{at}{T}\geq t_0\right\}=\frac{P\{t<T<t+dt\}}{P\{T\geq t_0\}}$$

$$\frac{a(t)dt}{1-A(t_0)}=\frac{\lambda e^{-\lambda t}dt}{e^{-\lambda t_0}}=\lambda e^{-\lambda(t-t_0)}dt \qquad (3)$$

Which shows that the entire inter-arrival time and the residues inter-arrival time after time t_0 have the same exponential distribution. Taking $t = t_0$, we see that the probability of arrival during $(t, t + dt)$does not depend upon the time since the last arrival. That is why Poisson arrivals are also called random arrivals. According to the basic structure, queues can either be Markovian or non-Markovian .

Arrival and Server: In order to model the arriving patients, we must consider the process that generates them. In this paper, arrivals are always stochastic and generated according to a Poisson process. Thus, by letting $k \in \boldsymbol{N_0}$ define the number of arrivals in a time-interval of size $\boldsymbol{t} \in \boldsymbol{R} \geq \boldsymbol{0}$ we have that

$$Prob\{k \text{ within } t\} = p_k(t) = e^{-\lambda t}\frac{(\lambda t)^k}{K!} \qquad (4)$$

$\lambda \epsilon \mathbb{R} > 0$ is the constant rate at which patients are arriving at the system.

Distributions

A queue of the type M/M/c can be viewed as a birth-death process. That is, a continuous-time Markov chain where the system can only change to the state $s \in \{0,1,2,\ldots,\infty\}$ through, the state of the process is either s = s + 1 (a single birth) or s = s – 1 (a single death) .

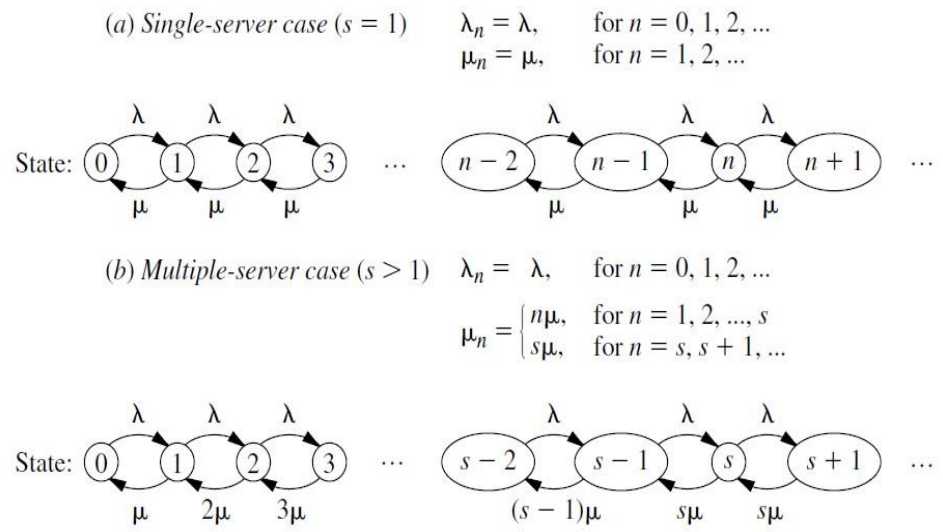

Figure 2: Infinite M/M/s model

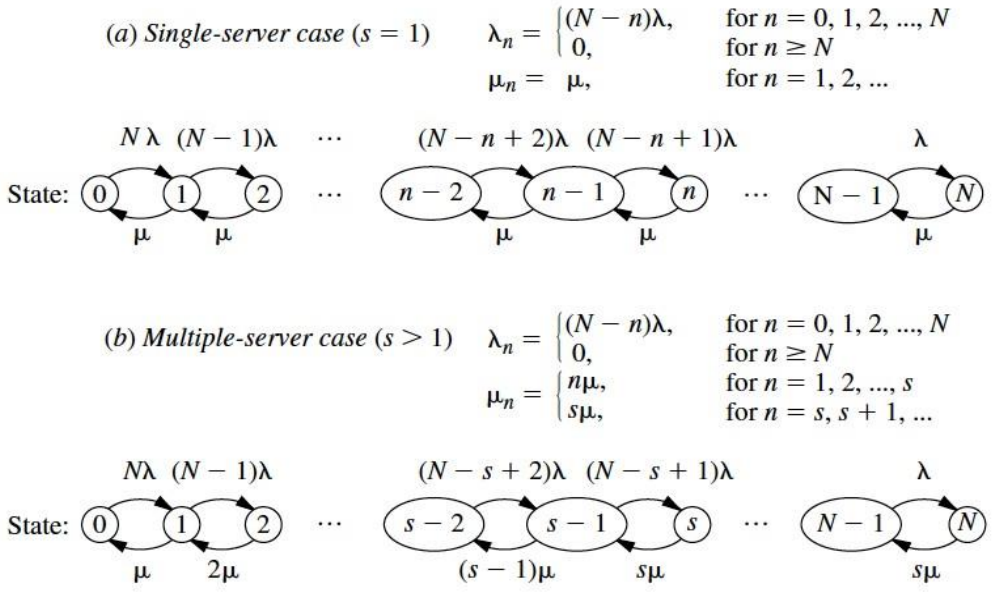

Figure 3: Finite M/M/s model

The time being considered in the process depicted in Figure 4, where s denotes the no. of patients in the system. This number increases with a rate corresponding to the arrival rate λ and decreases with the rate $\mu_S = s\mu$, where $\mu R>0$ is the rate at which patients are treated at each server (i.e. the reciprocal of the mean inter-service time). However, this only applies as long as $s \leq c$; otherwise, the rate is bounded at $\mu_S = c\mu$.

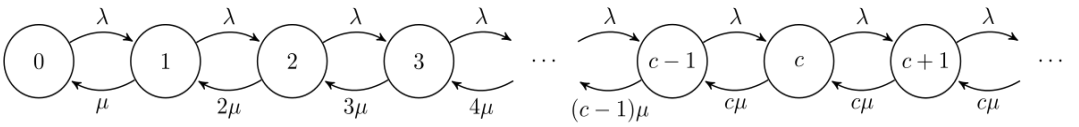

Figure 4: The state transitions associated with the $M/M/c$ queue.

A system is considered that has capacity-limit such that patients are lost from the system if all servers are occupied on arrival. That is, a queue of the type M/M/c/c, where the last c indicates capacity is equal to the no. of doctors/servers. In practice, this type of system occurs whenever patients do not wait for a resource (e.g. a nurse or a physician), but instead are relocated to a location where capacity is still available. Thus, the total no. of patients in the system is defined as s ∈ {0,1,...,c−1,c}, so the associated birth-death process comprises a finite number of states (Prasad, and Koka, 2015).

3.2 Queuing Models

Methodologically author mainly focus on evaluating the different instances of patient flow based on Markov chain modelling and using queue models. Design of queuing systems deals with determining the optimum system parameters such as the service rate, number of server's, number of service facilities, desirable waiting space or the appropriate queue discipline. In service operations management, it is always recommended to use a standard methodology that enables to understand the current processes, determine the desired changes and improve them. The present paper aims at optimizing queues in OPD department of AIIMS. Using in-depth analysis of OPDs from different dimensions, the objectives involved in the study is to do a study on different departments/ OPDs at AIIMS, Delhi. The input parameters for queuing models M/M/1 and up to M/M/8: Poisson-exponential, multiple server model-infinite populations are calculated using number of patients in the system denoted by n, average arrival rate denoted by λ and average service rate per server denoted by μ.

Markov chains are stochastic processes that are based on the notion that a system can be defined by a set of states referred to as the state space, S. The process can only attain a single state S at a time, but in return change (or transition) between the states as the process evolves. Further, the name *Markov* is derived from the Markovian property of the process, meaning that if at time t_k the process is in state s_k, then a transition at time t_{k+1} to a new state s_{k+1} is only dependent on s_k, and not on the history of the system (Suganthi, C. and S., 2020). Say we let $X(t) \in S$ define a stochastic process that evolves $t \geq 0$, and further that this process attains the states in a sequence $t_0 < t_1 <$--------$< t_{k-1} < t_k < t_{k+1}$, where $k \in N_0$ is the index of the sequence, then for a Markov chain(Taha, 2017)

$$Prob\{X(t_{k+1}) = s_{k+1} \mid X(t_k) = s_k \\ X(t_{k-1}) = s_{k-1}, \ldots X(t_0) = s_0\} = Prob\{X(t_{k+1}) = s_{k+1} \mid X(t_k) = s_k\} \quad (5)$$

Consider, for instance, the birth-death process for the aforementioned M/M/c queue in Figure 4. The number of patients when looking at the health care system

is either due to a recent arrival or discharge. Furthermore, because the inter-arrival and inter-service times follow a continuous distribution, a transition can occur at any point in time; hence $t \in R_0$.

3.3 Optimization

Minimizing this function will ultimately lead to an increased patient waiting time since fewer resources will be available to serve the patients. For this reason, the department has identified a set of nodes in the patient's path, denoted J, where the waiting time should not exceed a specific limit for a fraction of the patients. The department has additionally identified a relationship between the number of each employed staff type and the fraction of patients that exceed this limit, which can be modelled by a waiting time function W_j (x) for all j∈ J. Additionally, function W_j (x) is constrained by a lower bound denoted bj for all j∈ J . Further, assume that all staff types are subject to several departmental rules, union settlements, and practical limitations which can be modelled by the system of linear inequalities Ax≥ β, where β is a vector of length |K|, A is a |K|×|I| matrix. K is a set that comprises all necessary staff-constraints. For the convenience of this example, we assume that these constraints can be expressed by employing x directly without introducing any further dimensions to the vector, with this finally yielding the optimization problem.

Minimize cx
Subject to, Wj (x) ≥ bj where, ∀j ∈J, Ax ≥ β and x∈N_0

Due to the complexity of the waiting time functions, W_j (x)j∈J, we assume that optimality cannot be proven for all the constraints by employing any known solution approach (Taylor and Karlin, 2014).

3.4 Model Formulation

The M/M/1 model was designed considering a single-server system that sees K classes of patients. Their arrival rates differentiate the patients, processing requirements, mean arrival rate (λ), Service rate (μ) and (potentially) long waiting times. The long waiting times can alternatively be viewed as weights (high patient load, time of arrival, age, few doctors, day of arrival, treatment related diagnosis, providers, record clerks and clinicians, etc.) of importance among patient classes, where a patient class with a higher weight is more critical in comparison to lower weight. We consider multiple settings in which some Patient classes are of higher priority than others. Here, the holding long waiting times among the less urgent patients allows flexibility in modelling the relative importance of the less urgent patients, while maintaining that they are lower priority than the high-priority patients. The state space is $\mathbb{X} := \mathbb{Z}_+^K$, where the k^{th} component ($k\epsilon[K] := \{1, \ldots, k\}$) of a state $x\epsilon$ X. For each state x, the action space is A(x):= $\{k \in [K]: x_k > 0\}$: the server may be allocated to work on any class that is present in the system. In the particular case that $x = 0\epsilon\mathbb{R}^K$, we set A(x) = $\{-1\}$, where −1 denotes a "dummy" action to represent (forced) idling. Note that with this definition of the action space$A := U_{x\epsilon X}$ A(x), we restrict ourselves to non-idling policies.

Letting $e_k \epsilon \mathbb{R}^K$ $(k\epsilon[K])$, we describe the transition dynamics of the system by the generator

$$G(y|x,a) = \begin{cases} \lambda_k & y = x + e_k \\ \mu_a & y = x - 1\{x_a > 0\} e_a \\ -\sum_{k=1}^{K} \lambda_{k-} \ \mu_a \ 1\{x_a > 0\} & y = x \end{cases}$$

Where $1\{\cdot\}$ denotes the indicator function. Waiting-Time function W_k: X → R+ is $W_{k(x)} = h_k x_k$ .In this setting class, we can restrict our search for an optimal policy to the class of *stationary policies* (Fiems & De Vuyst, 2018). A stationary policy assigns actions based only on the current state of the system, regardless of the point in time at which the decision is being made. More formally, a stationary policy σ is a collection of probability distributions (σ_x) $x \in X$, such that Markov chain $X^\sigma = \{X^\sigma(t): t \geq 0\}$ on the state space $\mathbb{X}$ (Taylor and Karlin, 2014). The order of Patient service given is based on their priority and within each priority class FCFS. In this entire paper, the patient service time is assumed to be exponential and is class dependent, and there is a single server (1) & multiple servers (8) in the system. The single server and multi-server variables are defined as:

I-Poisson-exponential, M/M/1 model (single server queuing model)

L_q = Expected number of patients waiting in the queue

Lq = $\frac{\rho^2}{1-\rho}$ Where ρ= λ/ μ, λ = arrival rate, μ = service rate

L_s = Expected number of patients in the system (waiting + being served)

$$Ls = \frac{\rho}{1-\rho}$$

W_q = Expected waiting time in the queue

$$W_q = \frac{\rho}{\mu-\rho}$$

W_s = Expected time a patient spends in the system

$$W_S = \frac{1}{\mu-\lambda}$$

II-Poisson-exponential, M/M/s model (multi server queuing system)

Lq = The average number of patients in line waiting for service is

$$L_q = \frac{\left(\frac{\lambda}{\mu}\right)^K}{K!^{(1-\rho)^2}} P_0$$ where K=Number of servers, ρ= λ/Kμ

Ls = Expected number of patients in the system (waiting + being served)

$$L_S = L_q + \frac{\lambda}{\mu}$$

Wq = Expected waiting time of a patient in the queue waiting for service is

$$W_q = \frac{L_q}{\lambda}$$

W_S = Expected time a patient spends in the system

$$W_S = Wq+1/\mu$$

4 Study Area

There are 19 single window operational counters in AIIMS, where patients could seek their next appointments and dates for the various tests that were necessary. A total of 52 counters are operational in public reception center for the registration of patients. The AIIMS runs general medical clinics daily and specialized clinics from Monday to Friday (AIIMS, 2020). Most patients are seen on a walk-in basis without scheduled appointments with few patients admitted for day observation and minor procedures. There are 22 counters for current booking and 18 are the fast tract counter which took less than a minute for the registration of patients.

4.1 Data collection

Detailed queuing secondary data was collected at the designated AIIMS Hospital departments in Delhi over three years (1st January 2015 to 31st July 2017), a few specialist OPDs and pharmacy. The primary queuing data was collected for a total sample of 1200 patients per doctor (1st July to 30th July 2020) (AIIMS, 2020). Some of the forms depicting incomplete information are not considered in the study. Therefore, the data of 1200 patients were considered for queuing analysis. The data was collected during the working days (Monday-Friday) and period: 8:30 am to 01:00 pm.

4.2 Study population

The study population included all patients seeking outpatient care services at the AIIMS, Delhi health services during the four weeks of study.

Sample size: In total, 1200 respondents were recruited during the four-week study period.

Variables: A fundamental principle of queuing theory is: L = λW (Little's law. (Little, 1961)**.** Average queue size/ No. of patients in the queue (L) = Mean arrival rate x the average patient waiting time (W).

The dependent variable: patient waiting time while the
Independent variables: were the demographic factors like age, sex, patient condition (employee, employee dependent and student) and patient arrival time

Waiting time: The difference between when a patient arrived and when service started

Service Time: Difference between the time a patient entered until when he/she exited the system/mean time in the system

Length of Queue: The length of the queue at a particular point in time (Prasad, V.H and Koka, 2015)

5 Results and Discussion

The study has been carried out at AIIMS, Delhi to minimize the queue length. The three years of data collected from various departments of the hospital are shown in figure 5. The bar graph shows the comparison between the number of patients who visited the hospital. In the year 2016, the maximum numbers of patients arrived at the hospital in all the departments (AIIMS, 2020). The comparison between the mean patients and new patients shows that new patients turn up is significantly large in comparison mean no. of patients (as shown in figure 5). The number of New Appointment Attended (NAA) at AIIMS found to be approximately 2200, 3300, 2100 in 2015, 2016 and 2017 respectively average daily flow. The significant increase in the arrival of the patient definitely would lead to the generation of a long queue. The efforts for the minimization of queue length for the best treatment in minimum time is the urgent requirement for all the hospitals.

Table 2: Mean weekly data of AIIMS general OPD at each server

Days	No. of patients	Average waiting time in queue (mins)	Average service time (min)	Total time in system (min)
Monday	45	162	23	185
Tuesday	41	155	18	173
Wednesday	39	160	16	176
Thursday	41	158	20	178
Friday	35	145	22	167
Average	40	156	20	176

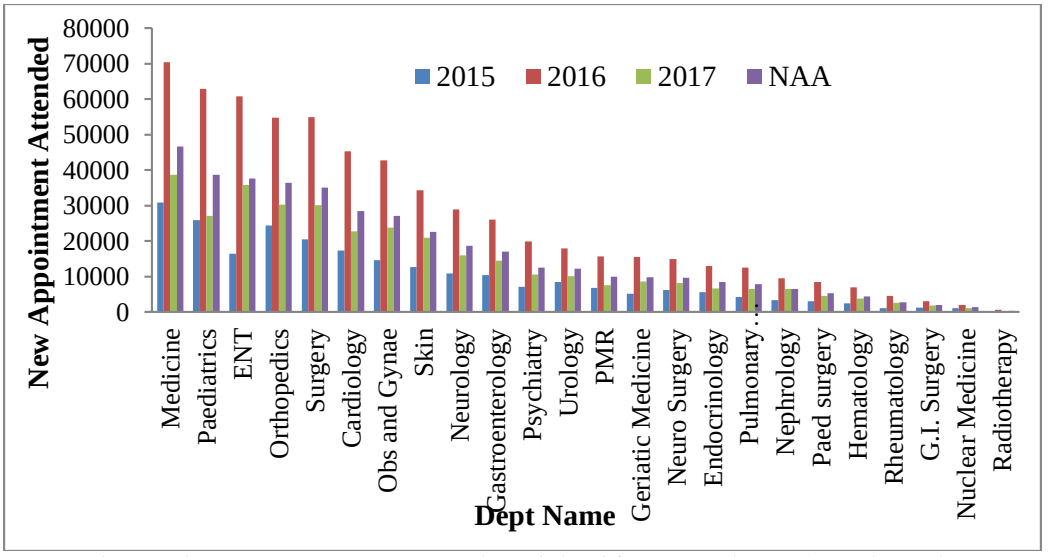

Figure 5: AIIMS Department wise visited/NAA patients (2015-2017)

AIIMS caters to a patient load of about 40- 50 patients per day in general OPD only, and the patients may be admitted in medical or surgical emergency (AIIMS, 2020). This strategy paid off for 14 per cent of patients who started arriving early at 6:00 am in the hospital. For the majority of patients in every department, then, the strategy of coming early did not pay off; in fact, it had the opposite effect of lengthening queues and increasing waiting time (Ryan, 1963). The patients in hospitals are managed by distributing of 8-10 according to the number of doctor's available. It was observed that out of the total people waiting outside the gates at 6 am, the patient count is approximately 50% of that, and the rest of the people are accompanied/escort to patients. All the patients are served on first come first served basis. However, it was found that maximum patients arrive in between 7:00am to 10:00am. It further noted that patients were not seen until 8.30 am. After adjusting for all the other factors, a fascinating phenomenon is seen for patients who arrived after 11:00. This may be due to breaking off times of some doctors. The distribution on patients on weekly basis is shown in table 2. The distribution shows that maximum patients arrive on Monday and least number of patient approach on Friday. The Figure 6 and 7 shows the average waiting time in queue and total time spent in system by the patients.

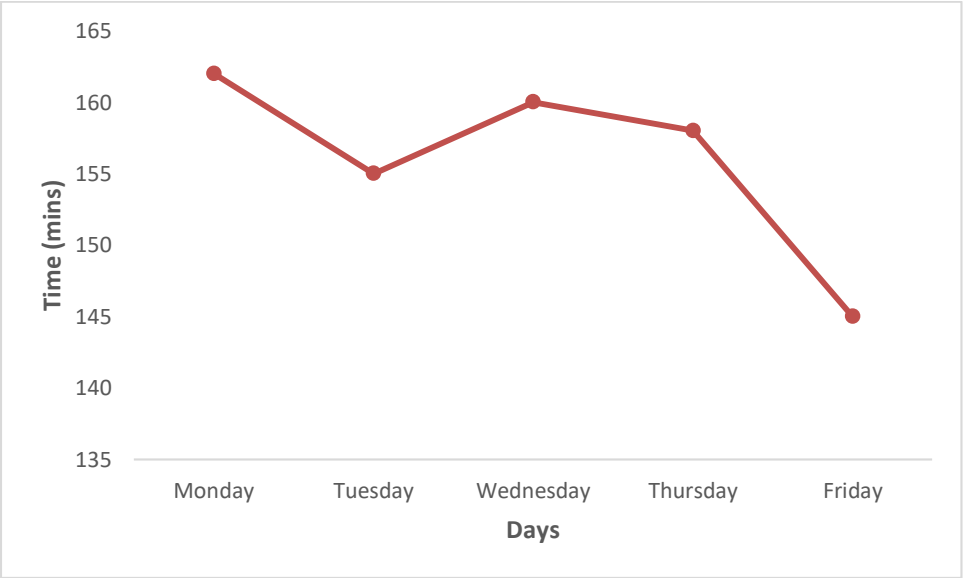

Figure 6: Average waiting time in queue of patients (in mins)

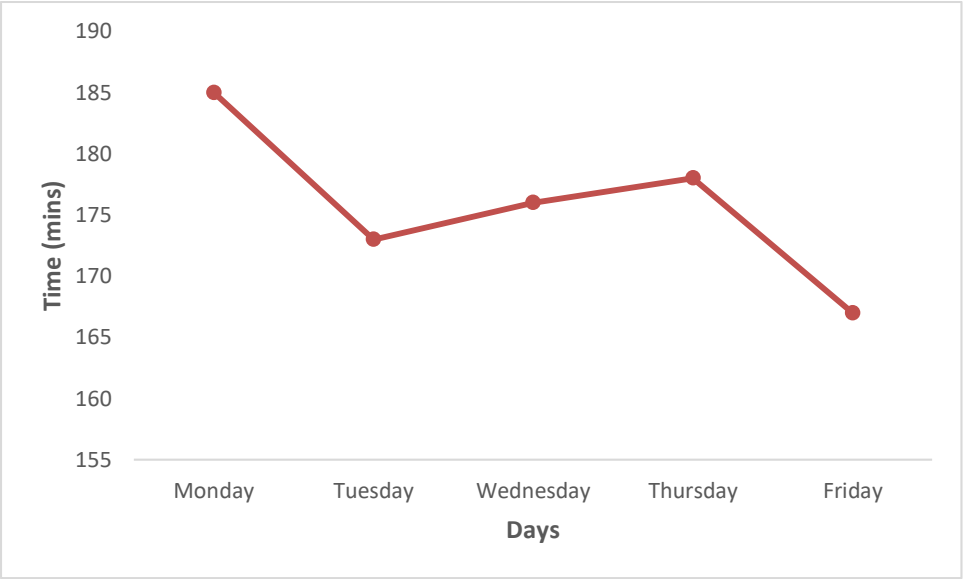

Figure 7: Total time spent by patient in hospital (in mins)

Queuing model analysis shows that the optimal service level where the waiting time is minimized, as shown in table 3. The assessment of the results demonstrates that some general relationships are underlying the waiting times. Research data analysis revealed an average of 200 minutes (i.e. 3.38 hours) of waiting time at AIIMS, Delhi, with the increase in the service levels/servers, average waiting times will be reduced while decreasing in the service level//servers induces increasing average waiting times. As expected, the waiting time of patient demands is directly proportional to the condition of the patients for whom a server/doctor/service provider is responsible for the change. Also, as the patient's load on the service provider decreases, the supplementary decrease in service delay is inversely proportional to the demand priority-that is, the higher the priority, the smaller the possible decrease in the delay of service. A study as a result of this noted that most doctors take their breaks between 1 pm-2 pm, resulting in a one-hour interval of minimal service. A similar application of the methodology also draws attention to several shortcomings of this type of models (Yaduvanshi et al., 2019). The author observed that online appointment system at AIIMS ends long queues, patients were able to take appointments over the phone (through the call center), IVRS (toll-free), on-site (Registration counters) and on the web (e-hospital), (AIIMS, 2020). Similarly, Queuing models M/M/1 to M/M/8: Poisson-exponential, multiple server model-infinite preemptive repeat-different and non-preemptive priority queues with thresholds on the maximum number of patients allowed to be in service and the queuing system (Wagner, 1997). In this paper, waiting for time distribution of a queuing process was obtained as similar the waiting time distribution of a set of particular events or patterns of an imbedded Markov chain which portrays the original queuing process (Chang, 2011). The author identified the research gap of queuing factors specifically for a queue system with no appointments in AIIMS Hospital. Most important here is the fact that the analytic solution of the model relies on the poison arrivals see time averages and Poisson similar nature of the arrival process.

Table 3: Estimation using Queuing Models

Characteristics	M/M/2	M/M/4	M/M/6	M/M/8
L	5.63	3.92	1.66	1.23
Lq	3.78	0.07	0.03	0.01
W (hr)	3.38	2.35	1.01	0.18
W_q (hr)	2.38	1.35	0.18	0.01
ρ	0.83	0.42	0.28	0.21

6 Conclusions

The present study specifically targeted the evaluation of a queue system with prior appointments in an outpatients' department of a busy Government. Hospital (AIIMS Delhi) in India, and has identified the operational characteristics of different queuing variables particularly within this context. Evaluated the queue system of an outpatients' department in a busy AIIMS hospital. In this paper, the author considers only queue models having Patients arrive at the AIIMS as Poisson processes and the service- time distribution being exponential. Research data analysis revealed an average of 4 hours of waiting time at AIIMS. Open Waiting time include the queue model of time-dependent demand-side rates, increases patient load as waiting times increases, high degree of variability in service times.

AIIMS hospital often uses flexible working schedules for overtime, variable server/doctor capacity. In this paper, fundamental quantities, such as the (effective) mean AIIMS hospitalization time and the patient's load, become functions of the queuing model's primitives. The author, therefore, begins by characterizing Lq, Ls and Wq, Ws maximal throughput. Waiting time may be reduced by increasing more server/ doctors. On the demand side/arrival rates as a response to queue lengths/Waiting timeless than service side time spent by the servers or doctors= high levels of Patients satisfaction. On the demand side or waiting time greater than service side time spent by the servers or doctors that means high levels of patient's dissatisfaction. Tradeoffs: demand-side or waiting time is equal to service side time spent by the servers or doctors. Future research may construct to find out all the problems of queuing and minimization of queues, maximize operational efficiency and maximize patient's service productivity of queuing system of the outpatient service at AIIMS, Delhi.

References

AIIMS, C. (2020). Online OPD Appointment. AIIMS NEW. Retrieved 10th August 2020, from https://www.aiims.edu/en/component/content/article/79-about-aiims/9211-online-opd-appointment.html.

Akilan Arunkumar, A. (2011). Delays and Waiting - A Challenge for Hospitals' Analysis of Outpatient Service using Queuing Model. Prabandhan: Indian Journal of Management, 4(4), 23.

Al-Begain, K., Heindl, A., & Telek, M. Analytical and Stochastic Modeling Techniques and Applications.

Bahadori, M., Mohammadnejhad, S., Ravangard, R., & Teymourzadeh, E. (2014). Using Queuing Theory and Simulation Model to Optimize Hospital Pharmacy

Performance. Iranian Red Crescent Medical Journal, 16(3).
Bako, I., Utoo, P., & Ikughur, J. (2017). Improving the Efficiency of Outpatient Services at Benue State University Teaching Hospital using the Queuing Theory. International Journal Of Statistics In Medical Research, 79-83.
Barlow, G. L. (2002). Auditing hospital queuing. Managerial Auditing Journal.
Chang, F. (2019). Optimization analysis of management operation for a server farm. Quality Technology & Quantitative Management, 17(3), 307-318.
Chang, H.-M. (2011). Waiting-Line Problems with Priority Assignment, and its Application on Hospital Emergency Department Wait-Time. 127-128.
Dai, J., & Shi, P. (2014). A Two-Time-Scale Approach to Time-Varying Queues for Hospital Inpatient Flow Management. SSRN Electronic Journal.
F. S. Hillier and G. J. Lieberman, "Queuing Theory", in Introduction to Operations Research, 5th ed, McGraw-Hill, Inc, U.S., 1990.
Fiems, D., & De Vuyst, S. (2018). From Exhaustive Vacation Queues to Preemptive Priority Queues with General Interarrival Times. International Journal of Applied Mathematics And Computer Science, 28(4), 695-704.
Hallal, L. (2015). Queuing models for long term care wait time reduction & capacity optimization. Saint Mary's University.
https://www.aiims.edu/en/component/content/article/79-about-aiims/9211-online-opd-appointment.html.
Ibrahim, I. M., Liong, C. Y., Bakar, S. A., Ahmad, N., & Najmuddin, A. F. (2017, April). Minimizing patient waiting time in emergency department of public hospital using simulation optimization approach. In AIP Conference Proceedings (Vol. 1830, No. 1, p. 060005). AIP Publishing LLC.
Joseph, J. (2020). Queuing Theory and Modeling Emergency Department Resource Utilization. Emergency Medicine Clinics of North America, 38(3), 563-572.
Kalashnikov, V. (1990). Regenerative queuing processes and their qualitative and quantitative analysis. Queuing Systems, 6(1), 113-136.
Kendall, D. (1953). Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. The Annals of Mathematical Statistics, 24(3), 338-354.
Kirpichnikov, A., & Titovtsev, A. (2016). Physical and Mathematical Queues in The Applied Queuing Theory. International Journal of Pure and applied Mathematics, 108(2).
Lee, H., Chung, D., Lee, S., & Chae, K. (1997). Server unavailability reduces mean waiting time in some batch service queuing systems. Computers & Operations Research, 24(6), 559-567.
Lefeber, E., Lämmer, S., & Rooda, J. (2011). Optimal control of a deterministic multiclass queuing system for which several queues can be served simultaneously. Systems & Control Letters, 60(7), 524-529.
Legros, B. (2018). Waiting time based routing policies to parallel queues with percentiles objectives. Operations Research Letters, 46(3), 356-361.
Leonardi, E. (2014). Throughput optimal scheduling policies in networks of constrained queues. Queuing Systems, 78(3), 197-223.
Lin, C., & Wu, C. (2012). Mathematically modelling the effects of pacing, finger strategies and urgency on numerical typing performance with queuing network model human processor. Ergonomics, 55(10), 1180-1204.
Little, J. (1961). A Proof for the Queuing Formula: L= λW. Operations Research, 9(3), 383-387.

Mittal, H., and Sharma, N. (2020a). Assessment of Delhi NCR Traffic Using Queuing Model. International Journal of Advanced Science and Technology, 29(7), 5418-5424.

Mittal, H., and Sharma, N. (2020b). A Probabilistic Model for the Assessment of Queuing Time of Coronavirus Disease (COVID-19) Patients using Queuing Model. International Journal of Advanced Research in Engineering and Technology, 11(8), 22-31.

Mittal, H., & Sharma, N., (2022a). Modeling of Communication Network with Queuing Theory under Fuzzy Environment. Mathematical Statistician and Engineering Applications, 71(2), 122-137.

Mittal, H., & Sharma, N., (2022b). Operational Optimization of Toll Plaza Queue Length Using Microscopic Simulation VISSIM Model. Journal of Algebraic Statistics, 13(1), 418-425.

Nuyens, M. (2004). The Maximum Queue Length for Heavy-Tailed Service Times in the M/G/1 FB Queue. Queuing Systems, 47(1/2), 107-116.

Obulor, R., & Eke, B. O. (2016). Outpatient queuing model development for hospital appointment system. International Journal of Scientific Engineering and Applied Science (IJSEAS), 2(4), 15-22.

ORS Patient Portal. Ors.gov.in. (2020). Retrieved 10th August 2020, from https://ors.gov.in/copp/cancel_tc.jsp?NICSecurityORS=R1SA-TL1D-1X67-24KB-MP7E-WV4V-MNCX-0LFO.

Prasad, V., V.H, B., & Koka, T. (2015). Mathematical analysis of single queue multi server and multi queue multi server queuing models: comparison study. Global Journal of Mathematical Analysis, 3(3), 97.

Rotich, T. (2016). Utility Analysis of an Emergency Medical Service Model Using Queuing Theory. British Journal of Mathematics & Computer Science, 19(1), 1-18.

Setiawan, A., & Nugraheny, D. (2019). Mobile-Based Outpatient Queue System Using the Priority Scheduling And First Come First Served Scheduling Method. Angkasa: Jurnal Ilmiah Bidang Teknologi, 11(1), 54.

Sharma, N., Mishra, GD., and Kumar A (2011). Optimization Model Of Toll Plaza Using A Combination Of Queueing And Simulation. International Journal of Mathematical Sciences & Engineering Applications, 5(VI), 331-346.

Kumar, S. (2022). A quest for sustainium (sustainability Premium): review of sustainable bonds. Academy of Accounting and Financial Studies Journal, Vol. 26, no.2, pp. 1-18

Allugunti V.R (2022). A machine learning model for skin disease classification using convolution neural network. International Journal of Computing, Programming and Database Management 3(1), 141-147

Sharma, N., Mishra, GD., and Kumar A (2011). Simulation of Service Station of Vehicles With One Server. International Journal of Computer, Mathematical Sciences and Application, 4(1-2), 85-97.

Suganthi, M., C., A., & S., A. (2020). Mathematical modeling in comparative relation of single queue multi server and multi queue multi server queuing models. Malaya Journal Of Matematik, S(1), 331-334.

Tadj, L. (1996). Waiting in line [queuing theory]. IEEE Potentials, 14(5), 11-13.

Taha, H. (2017). Operations research. Pearson.

Dwijaya, A., & Atmaja, M. H. S. (2022). Clinical and imaging findings of klippel-trenaunay syndrome: A case report. International Journal of Health & Medical Sciences, 5(1), 145-149. https://doi.org/10.21744/ijhms.v5n1.1860

Taylor, H., & Karlin, S. (2014). An Introduction to Stochastic Modeling. Elsevier Science.

Thomas, R. (2013). Application of Queuing Analytic Theory to Decrease Waiting Times in Emergency Department: A Review. Archives of Trauma Research, 2(1), 54-5.

Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). The COVID-19 pandemic. International Journal of Health Sciences, 5(2), vi-ix. https://doi.org/10.53730/ijhs.v5n2.2937

Titovtsev, A. (2016). The Concept of Higher Orders Queues in The Queuing Theory. International Journal of Pure and Applied Mathematics, 109(2).

Wagner, D. (1997). Waiting times of a finite-capacity multi-server model with non-preemptive priorities. European Journal of Operational Research, 102(1), 227-241.

Yaduvanshi, D., Sharma, A., & More, P. (2019). Application of Queuing Theory to Optimize Waiting-Time in Hospital Operations. Operations and Supply Chain Management: An International Journal, 165-174.

Zhand, H., Shi, D., & Hou, Z. (2014). Explicit Solution for Queue Length Distribution of M/T-Sph/1 Queue. Asia-Pacific Journal of Operational Research, 31(01), 1450001.