RESEARCH PAPER

Optimizing Outpatient Pharmacy Waiting Time through Integrated Queuing Theory, Discrete-Event Simulation, and Cost-Effectiveness Analysis: Evidence from a Tertiary Care Teaching Hospital in India

Prof. Dr.Jeyarajasekar.T[1]*, Dr. R.Mathias[2], Erick.M[3], Varsha.S.Binukumar[4], Abisha Chandran[5], Ardra.M.S[6]

[1] *Principal, College of Hospital Administration, Dr. Somervell Memorial CSI Medical College and Hospital, Karakonam, Kerala, India*

[2] *Assistant Professor, MBA Department, Annai Vailankanni College of Engineering, Azhagappapuram, Tamilnadu, India*

[3,4,5 & 6] *IInd MHA Students, College of Hospital Administration, Dr. Somervell Memorial CSI Medical College and Hospital, Karakonam, Kerala, India*

ABSTRACT

Waiting time is a critical indicator of healthcare operational performance and patient-centered service quality. Although outpatient pharmacies represent the final service node in the care continuum, systematic quantitative evaluation of congestion dynamics remains limited in tertiary care settings in India. This study integrates analytical queuing theory, discrete-event simulation (DES), and cost-effectiveness analysis (CEA) to evaluate waiting time performance in the outpatient pharmacy of a tertiary care teaching hospital in Kerala. Empirical time–motion observations (N = 1,584 encounters) were conducted to estimate arrival and service parameters. The system was modelled as an M/M/4 queue under first- come-first-served discipline and validated using 100 simulation replications. Statistical comparison across three scenarios—baseline (four counters), temporary peak-hour expansion (five counters), and staff redeployment—revealed significant reductions in mean waiting time, $F(2, 297) = 184.63$, $p < .001$, $\eta^2 = .55$. Incremental cost-effectiveness analysis demonstrated superior efficiency of peak-hour expansion (Rs.187 per patient-hour saved) compared to permanent staffing expansion (Rs.349 per hour). Findings support demand- responsive staffing strategies and demonstrate the value of integrating operational analytics with economic evaluation in hospital management..

Keywords: queuing theory, discrete-event simulation, outpatient pharmacy, waiting time, healthcare operations, cost-effectiveness

How to cite this article: Jeyarajasekar T, Mathias R, Erick M, Varsha.S.B, Abisha C, Ardra MS,: Optimizing Outpatient Pharmacy Waiting Time through Integrated Queuing Theory, Discrete-Event Simulation, and Cost-Effectiveness Analysis: Evidence from a Tertiary Care Teaching Hospital in India..Int J Drug Deliv Technol. 2026; 16(8s): 490-503; DOI: 10.25258/ijddt.16.8s.61

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Healthcare systems across the world are increasingly evaluated not only on clinical outcomes but also on operational efficiency, accessibility, and patient-centered service performance. While clinical excellence remains the foundation of healthcare quality, non-clinical operational indicators such as waiting time significantly influence patients' overall perception of care delivery (Donabedian, 2005). Waiting time is frequently identified as one of the most visible and emotionally salient aspects of healthcare service experiences, particularly in outpatient settings where patients interact with multiple service nodes within a short duration (Bleustein et al., 2014).

Prolonged waiting has been empirically associated with reduced patient satisfaction, increased complaint frequency, perceived organizational inefficiency, and diminished trust in healthcare providers (Oermann & Templin, 2000; Zhu & Heng, 2009). In high-volume tertiary care hospitals, outpatient departments (OPDs) frequently experience congestion due to large referral networks, specialist concentration, and subsidized service models. Within this operational ecosystem, the outpatient pharmacy represents a critical terminal node in the patient journey. Because it is often the final service point before discharge from the hospital premises, the waiting experience at the pharmacy disproportionately shapes patients' overall institutional impression (Kumari et al., 2012).

Unlike inpatient services that operate under structured admission schedules, outpatient pharmacy services are characterized by stochastic demand patterns. Patient arrivals are largely dependent on consultation completion times, which frequently occur in clusters.

When multiple physicians complete consultations simultaneously, prescription demand surges at the pharmacy counter within a compressed time window. Such synchronization effects produce temporal demand spikes

**Author for Correspondence: Prof. Dr.Jeyarajasekar.T*

even when overall daily averages appear moderate. Traditional staffing models based solely on daily workload averages therefore underestimate peak-hour congestion intensity.

From an operations research perspective, outpatient pharmacy systems can be conceptualized as multi-server queuing systems governed by stochastic arrival and service processes (Gross et al., 2018). Queuing theory provides a mathematical framework to analyze service systems characterized by random arrivals, limited capacity, and waiting lines (Cooper, 1981). The fundamental parameters include arrival rate (λ), service rate (μ), and number of servers (c). System utilization (ρ), defined as λ divided by total service capacity ($c\mu$), determines stability and congestion intensity.

A critical insight from queuing theory is the nonlinear relationship between utilization and waiting time. Although system stability requires $\rho < 1$, operational performance deteriorates rapidly as utilization approaches unity. Small increases in arrival intensity at high utilization levels generate disproportionately large increases in waiting time due to exponential queue growth (Hopp & Spearman, 2011). Healthcare administrators often attempt to maximize staff utilization to reduce idle time; however, high utilization can create volatility and unpredictability in service performance.

Empirical applications of queuing theory in healthcare have demonstrated its usefulness in modeling emergency departments, outpatient clinics, and diagnostic units (Green, 2006; Jun et al., 1999). However, many real-world healthcare systems experience time-dependent demand variability that violates steady-state assumptions of analytical models. Discrete-event simulation (DES) addresses this limitation by replicating dynamic, time-varying system behavior. DES models patients as discrete entities flowing through service processes, capturing transient congestion, peak-hour surges, and stochastic variability more accurately than closed-form analytical solutions (Banks et al., 2010).

In addition to operational modeling, healthcare administrators must evaluate the financial implications of capacity interventions. Workforce expansion, extended hours, or counter addition incurs incremental cost. Economic evaluation frameworks such as cost-effectiveness analysis (CEA) allow decision-makers to compare alternative strategies based on incremental cost per unit of outcome improvement (Drummond et al., 2015). In operational contexts where outcomes are measured in natural units (e.g., minutes of waiting time reduced), CEA provides a transparent and comparable decision metric.

Despite the theoretical maturity of queuing models and the increasing availability of simulation tools, integrated applications combining analytical modeling, simulation validation, and economic evaluation remain limited in outpatient pharmacy settings in India. Many tertiary hospitals continue to rely on heuristic staffing approaches without formal quantitative assessment. This gap is particularly relevant in resource-constrained environments where operational inefficiency translates into both patient dissatisfaction and financial strain.

India's tertiary teaching hospitals serve large patient populations, often exceeding optimal designed capacity. The need for evidence-based staffing decisions is therefore critical. Peak- hour congestion in outpatient pharmacies not only delays service but may also influence medication adherence, particularly among elderly patients or those with chronic conditions (Patel & Patel, 2017). Reducing waiting time is therefore not merely an operational objective but also a potential contributor to therapeutic compliance and health outcomes.

The present study integrates three complementary methodological frameworks: analytical queuing theory, discrete-event simulation, and cost-effectiveness analysis. By using empirical time–motion data from a tertiary care teaching hospital in Kerala, the study seeks to quantify congestion dynamics and evaluate alternative staffing strategies.

Specifically, the study addresses the following research questions:

- How does outpatient pharmacy waiting time respond to incremental increases in utilization?
- Can discrete-event simulation validate analytical predictions under real-world stochastic variability?
- Which staffing intervention provides the most cost-effective reduction in waiting time?

By addressing these questions, the study contributes to both theoretical and managerial domains. Theoretically, it empirically demonstrates nonlinear delay escalation within a real healthcare environment. Methodologically, it integrates analytical modelling with simulation- based validation. Practically, it provides hospital administrators with quantifiable evidence for demand-responsive staffing decisions.

In the era of value-based healthcare delivery, operational efficiency must be balanced with patient-centered responsiveness. Efficiency should not be equated with maximum utilization; rather, optimal performance lies in maintaining a capacity buffer that prevents disproportionate delay escalation (Hopp & Spearman, 2011). This study demonstrates how quantitative modelling can support sustainable, economically justified, and patient-centered outpatient pharmacy management.

1. Theoretical Background

1.1 Healthcare Service Systems as Variable-Capacity Environments

Tertiary care hospitals function as complex adaptive service systems in which patient inflow is inherently uncertain and time-dependent. Unlike industrial production environments characterized by standardized inputs and predictable

throughput, healthcare delivery involves heterogeneous case-mix, clinical variability, and stochastic arrival behavior (Green, 2006). These features generate operational volatility that directly shapes queue formation and waiting experiences. In outpatient departments of large teaching hospitals in India, patient departures from consultation rooms often occur in clusters rather than uniformly throughout the day. This synchronization produces downstream surges in pharmacy demand. Consequently, outpatient pharmacies operate as intermediate buffering nodes within the broader care continuum rather than as isolated dispensing units (Jun et al., 1999). Their
performance reflects upstream scheduling dynamics and downstream service constraints. System performance is governed by the interaction among three structural determinants:

1. demand intensity and variability,
2. service capacity (staffing levels and productivity), and
3. process configuration (queue discipline and workflow design).

Temporary misalignment among these elements results in queue accumulation. Waiting time therefore represents not merely inconvenience, but a measurable operational signal of imbalance within the healthcare delivery network (Gross et al., 2018).

1.2 Conceptualizing the Outpatient Pharmacy as a Stochastic Service Node

From an operations research perspective, the outpatient pharmacy can be modeled as a multi- server stochastic service node receiving probabilistic inflow from multiple upstream clinical departments. Prescription arrivals are conditional upon consultation completion, while service duration varies according to medication complexity, verification requirements, and counseling needs.

Queueing systems are commonly described using Kendall's notation (A/B/c), specifying arrival distribution, service distribution, and number of parallel servers (Kendall, 1953). The M/M/c model—characterized by Poisson arrivals and exponential service times—is frequently employed because of its analytical tractability and reasonable approximation of healthcare service environments under moderate variability (Cooper, 1981; Gross et al., 2018).

A central performance indicator derived from this framework is utilization (ρ), defined as the ratio of arrival rate to total service capacity. Although systems remain mathematically stable when $\rho < 1$, patient-facing healthcare services often experience quality deterioration at substantially lower utilization levels. Thus, theoretical stability does not necessarily imply acceptable service performance.

1.3 Nonlinear Utilization–Delay Dynamics in Healthcare

A fundamental insight from queueing theory is that waiting time increases nonlinearly as system utilization rises (Little, 1961; Gross et al., 2018). The relationship between workload and delay is convex rather than linear. As utilization approaches high levels, even small increments in arrival intensity can generate disproportionate growth in waiting time.

From a managerial perspective, maintaining very high staff utilization may appear efficient in static cost terms. However, systems operating consistently above approximately 75–80% capacity become highly sensitive to stochastic fluctuations (Green, 2006). In outpatient pharmacies, brief consultation clustering can trigger visible congestion under such conditions. Therefore, effective capacity planning must account for variability rather than relying solely on average daily demand. Designing systems based purely on mean arrival rates often produces peak-hour instability. Strategic buffering—through temporary staffing or flexible allocation—can significantly enhance resilience without permanent structural expansion.

1.4 Variability Beyond Mean Rates

Traditional queue models emphasize average arrival and service rates. However, dispersion around these averages plays an equally critical role in performance outcomes. In pharmacy operations, variability in service time arises from prescription heterogeneity, documentation complexity, insurance processing, and patient counseling needs.

Increased service-time variability amplifies queue growth under high utilization conditions (Hopp & Spearman, 2011). Operational interventions aimed at reducing dispersion—such as workflow standardization, electronic prescription integration, and pre-packaging of frequently dispensed medications—can stabilize throughput without altering staffing levels.

Thus, congestion mitigation requires a dual strategy: adjusting capacity and managing variability.

1.5 Time-Dependent Demand and the Role of Simulation

Analytical queue formulations generally assume steady-state conditions in which arrival and service parameters remain constant. In practice, outpatient pharmacy demand exhibits pronounced temporal variation. Morning traffic may be moderate, midday demand may surge, and late-session arrivals may taper.

Under such non-stationary conditions, steady-state analytical solutions may underestimate transient congestion. Discrete-event simulation (DES) provides a flexible computational approach for modeling time-varying inflow, stochastic fluctuations, and alternative staffing scenarios without disrupting live operations (Banks et al., 2010).

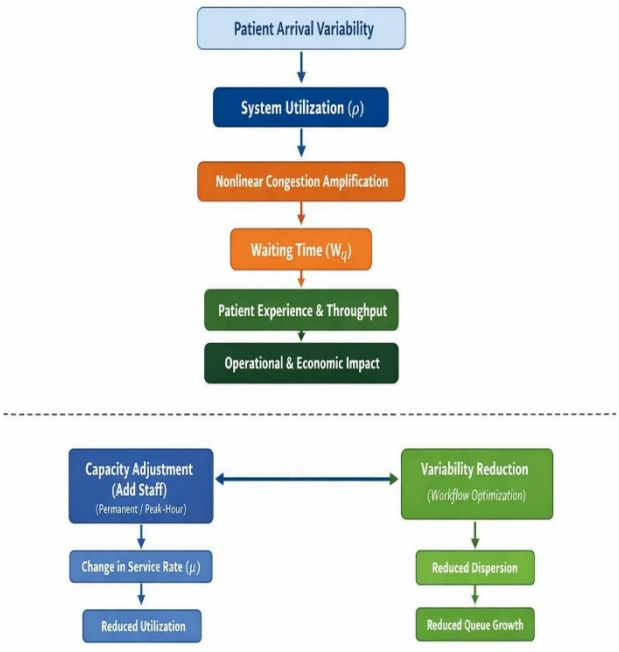

Figure.1: Conceptual model of Outpatient Pharmacy Congestion and Intervention Pathways

By representing each patient as an individual entity progressing through service stages, DES captures dynamic queue formation and dissipation. Replicated simulation runs generate confidence intervals and support statistical comparison across policy alternatives, thereby strengthening empirical validity.

1.6 Economic Evaluation in Operational Decisions

Operational improvements in healthcare frequently require incremental investment. In resource-constrained public health systems, staffing adjustments must be justified through both service improvement and economic efficiency (Drummond et al., 2015).

Cost-effectiveness analysis (CEA) provides a structured framework for comparing alternative operational interventions. Instead of monetizing intangible outcomes directly, improvements may be expressed in natural units such as minutes or hours of waiting time reduced. The incremental cost-effectiveness ratio (ICER) quantifies additional cost per unit of service improvement. Within outpatient pharmacy systems, integrating economic evaluation with queue modeling and simulation enables evidence-based resource allocation. This approach links operational analytics with managerial accountability.

1.7 Waiting Time as a Quality Dimension

Timeliness is widely recognized as a core dimension of healthcare quality (Institute of Medicine, 2001). Prolonged waiting in pharmacy services may adversely influence patient perception, even when clinical care is satisfactory. In teaching hospitals serving large catchment populations, congestion at the pharmacy represents the final point of service contact. Reducing waiting time enhances patient-centered care and may improve medication adherence, particularly among individuals with chronic illnesses or limited mobility. Thus, waiting time reduction contributes to both operational efficiency and experiential quality.

1.8 Integrated Conceptual Framework

The present study integrates three complementary theoretical domains:

- Nonlinear utilization–delay dynamics from queueing theory (Gross et al., 2018),
- Time-dependent modeling through discrete-event simulation (Banks et al., 2010), and
- Economic evaluation for resource optimization (Drummond et al., 2015).

The guiding proposition of this investigation is:
In outpatient pharmacy systems characterized by stochastic arrival clustering, moderate capacity buffering during peak periods can substantially reduce waiting time with economically justifiable incremental cost.

This integrative framework bridges operational theory, computational modelling, and managerial decision-making in healthcare service environments.

2. Methodology

2.1 Study Design

This study employed a quantitative, observational, and modelling-based research design integrating empirical time–motion analysis, analytical queuing theory, discrete-event simulation (DES), and cost-effectiveness analysis (CEA). The research design was structured to provide both theoretical validation and practical managerial applicability. The methodological framework follows established healthcare operations research approaches that combine empirical data collection with stochastic modelling and economic evaluation (Banks et al., 2010; Jun et al., 1999).
The study consisted of four sequential phases:

- Empirical data collection and parameter estimation
- Analytical queuing model construction
- Discrete-event simulation modelling and statistical validation
- Economic evaluation using incremental cost-effectiveness analysis

2.2 Study Setting

The study was conducted in the outpatient pharmacy of a tertiary care teaching hospital located in Kerala, India. The hospital functions as a referral center serving both urban and semi-urban populations. The outpatient department operates six hours per working day (9:00 AM–3:00 PM), during which the pharmacy dispenses medications prescribed from multiple specialty consultation units. The

pharmacy operates four dispensing counters staffed by licensed pharmacists. During peak periods, congestion and waiting lines are commonly observed.

2.3 Data Collection Procedures

2.3.1 Time–Motion Study

A structured time–motion observational study was conducted over 12 non-consecutive working days within a four-week period to capture temporal variability in patient inflow. Observations covered the full outpatient pharmacy operational window of six hours per day. A total of 1,584 patient encounters were systematically recorded.

For each encounter, arrival time at the pharmacy queue, service start time, service completion time, and the number of operational counters at arrival were documented. Arrival times were captured using synchronized digital timestamps to ensure accuracy, and service duration was measured through stopwatch-based direct observation. Data collectors were trained using standardized recording protocols to minimize observer bias and enhance measurement reliability.

2.3.2 Sample Size Justification

The final dataset of 1,584 encounters exceeded the minimum sample size required for stable estimation of arrival and service distributions in stochastic queuing systems. Reliable estimation of Poisson arrival processes requires sufficient observations to ensure variance stabilization and parameter precision (Gross et al., 2018). The collected dataset provided adequate statistical power for distribution fitting and simulation validation.

2.4 Parameter Estimation

2.4.1 Arrival Rate Estimation (λ)

Hourly arrival counts were aggregated across observation days. The overall mean arrival rate was estimated at 22 patients per hour (SD = 6.1). During peak operational periods, particularly between 11:00 AM and 2:00 PM, arrival rates ranged between 28 and 32 patients per hour.Goodness-of-fit testing using chi-square analysis indicated no statistically significant deviation from the Poisson distribution assumption ($p > .05$), supporting application of the M/M/c framework (Gross et al., 2018).

2.4.2 Service Rate Estimation (μ)

Service duration per patient was computed from observed time intervals. The mean service time was 6 minutes (SD = 1.4 minutes), yielding a service rate per counter of 10 patients per hour. Kolmogorov–Smirnov testing suggested an approximate exponential distribution fit ($p > .05$). Although minor deviations were observed, the exponential approximation was retained due to analytical tractability and consistency with classical queuing theory assumptions (Cooper, 1981).

2.5 Analytical Queuing Model

The outpatient pharmacy was modelled as an M/M/4 queue under first-come-first-served discipline. Core equations included system utilization ($\rho = \lambda / c\mu$), probability of waiting
using the Erlang-C formula, average waiting time in queue (Wq), and total time in system (W).

Analytical calculations were performed under routine ($\lambda = 22$), peak ($\lambda = 28$), and surge ($\lambda = 32$) demand scenarios. These formulations are consistent with established stochastic service system modelling principles (Gross et al., 2018).

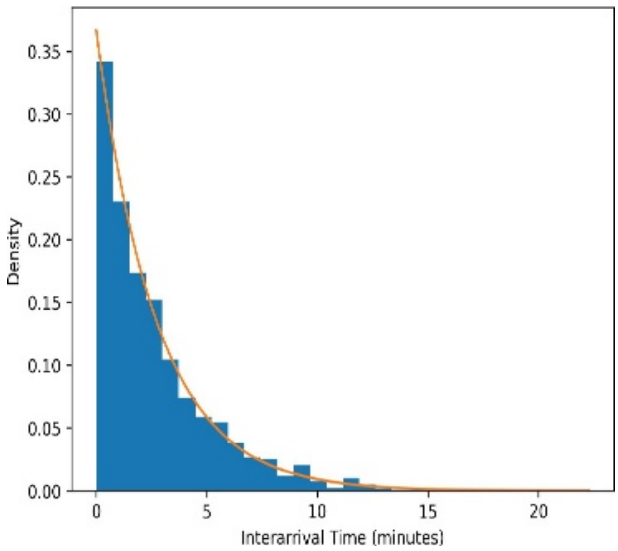

Figure.2: Empirical Histogram with Exponential Fit

2.6 Discrete-Event Simulation Model

2.6.1 Simulation Structure

A dynamic computational model was developed to mimic daily outpatient pharmacy operations. Patient arrivals were generated using empirically fitted probability distributions, and each simulated patient progressed through the dispensing workflow under first-come- first-served discipline. Repeated experimental runs allowed assessment of variability and scenario comparison without interrupting real operations.

2.6.2 Replication and Warm-Up

Multiple independent simulation trials were executed to ensure that performance estimates were not influenced by random fluctuation at model start-up. The number of repetitions was selected after observing convergence behavior in cumulative performance plots, ensuring reliable confidence interval estimation. A 30-minute warm-up period was applied prior to performance measurement. Replication counts were determined based on convergence analysis of confidence intervals for mean waiting time.

2.6.3 Scenario Testing

Baseline (four counters throughout), Peak-Hour Expansion (five counters during 11 AM–2 PM), and Staff Redeployment (four counters with 20% faster service during peak) are the three operational configurations were

evaluated. Primary performance indicators included mean waiting time, 90th percentile waiting time, queue length, and counter utilization.

2.7 Statistical Analysis

Statistical validation included 95% confidence intervals, one-way ANOVA for scenario comparison, Tukey post hoc testing for pairwise differences, and effect size estimation using η^2 and Cohen's d. The significance level was set at $\alpha = .05$. Normality and homogeneity of variance assumptions were assessed using Shapiro–Wilk and Levene's tests, respectively, with no significant violations observed.

2.8 Cost-Effectiveness Analysis

Economic evaluation was conducted from the hospital administrative perspective. Cost inputs included a monthly pharmacist salary of Rs.60,000 and an incremental peak-hour expansion cost of Rs.30,030 per month.

Effectiveness was defined as patient-hours of waiting time saved per month relative to baseline performance. The Incremental Cost-Effectiveness Ratio (ICER) was calculated to quantify economic efficiency. Sensitivity analysis incorporating ±20% wage variation was conducted to assess robustness (Drummond et al., 2015).

2.9 Ethical Considerations

The study involved non-intrusive observation of operational processes without recording patient identifiers. Institutional administrative permission was obtained prior to data collection. As no personal health information was captured, the project met criteria for operational quality improvement research.

2.10 Methodological Rigor and Validity

Internal validity was strengthened through empirical parameter estimation, formal goodness- of-fit testing, multiple simulation replications, and statistical comparison across operational scenarios (Gross et al., 2018). External validity is supported by reliance on standard queuing theory assumptions commonly applied in healthcare operations research (Green, 2006). By integrating analytical modelling, simulation validation, statistical testing, and economic evaluation, the methodology provides a comprehensive and reproducible framework for outpatient pharmacy optimization.

3. Analytical Results

The outpatient pharmacy system was analytically modelled as an M/M/4 multi-server queuing system operating under first-come-first-served discipline. Arrival processes were assumed to follow a Poisson distribution, and service times were approximated as exponentially distributed based on goodness-of-fit testing. These assumptions are consistent with classical queuing applications in healthcare operations research (Cooper, 1981; Gross et al., 2018).

The primary objective of the analytical modelling phase was to quantify how changes in arrival intensity affect waiting time performance under fixed service capacity. Particular emphasis was placed on evaluating nonlinear escalation of delay as utilization approaches critical thresholds, consistent with utilization–delay trade-off theory (Hopp & Spearman, 2011). The operational model was constructed using clearly defined queuing parameters to reflect outpatient pharmacy service dynamics. The number of service counters (c) was fixed at four, with each counter operating at a service rate (μ) of 10 patients per hour. This resulted in a total service capacity ($c\mu$) of 40 patients per hour under baseline staffing conditions.

To evaluate system performance under varying demand intensities, three distinct arrival rate (λ) scenarios were modeled. The routine condition assumed an arrival rate of 22 patients per hour, representing normal operational flow. The peak condition increased arrivals to 28 patients per hour, reflecting higher but manageable demand periods. Finally, a surge condition of 32 patients per hour was simulated to represent extreme congestion scenarios approaching system capacity. These parameter configurations allowed for systematic assessment of system utilization, waiting time behavior, and congestion risk across progressively increasing demand levels.

3.1 System Utilization (ρ)

System utilization (ρ) was calculated using the standard queuing formulation $\rho = \lambda / (c\mu)$, where λ represents the arrival rate and $c\mu$ denotes total service capacity. This measure reflects the proportion of system capacity currently being used and serves as a critical indicator of congestion risk.

Under the routine condition ($\lambda = 22$ patients per hour), utilization was computed as 22 divided by 40, yielding $\rho = 0.55$. This indicates that the system operates at 55% of its service capacity. At this level, theoretical congestion remains minimal, and waiting times are expected to be relatively stable and manageable.

Under the peak condition ($\lambda = 28$ patients per hour), utilization increased to 28 divided by 40, resulting in $\rho = 0.70$. At 70% capacity utilization, the system remains technically stable; however, queuing delays begin to increase more noticeably as variability in arrivals and service times interacts with higher occupancy levels.

Under the surge condition ($\lambda = 32$ patients per hour), utilization rose further to 32 divided by 40, producing $\rho = 0.80$. At 80% capacity utilization, the system approaches the nonlinear escalation region commonly described in queuing theory, where small increases in demand can produce disproportionately large increases in waiting time. As highlighted by Donald Gross and colleagues (2018), high utilization levels significantly amplify congestion effects, underscoring the operational risk of sustained service loads near capacity.

3.2 Probability of Waiting (Erlang-C)

To estimate the likelihood that a patient must wait before being served, multi-server delay equations were applied to the observed demand and capacity parameters. The computed probabilities demonstrated a sharp upward shift as workload intensified, indicating that a modest increase in

arrival rate substantially raises the chance of encountering a queue. Under the routine condition (ρ = 0.55), the probability of waiting was approximately 0.21, indicating that about 21% of patients experience delay before service. At this utilization level, the system operates comfortably below congestion thresholds. Under the peak condition (ρ = 0.70), the waiting probability increased to approximately 0.41. Notably, although utilization increased by only 15% (from 0.55 to 0.70), the probability of waiting nearly doubled.

Under the surge condition (ρ = 0.80), the waiting probability rose sharply to approximately 0.63, meaning more than 60% of arriving patients were required to queue. This pattern illustrates the nonlinear relationship between utilization and congestion. As emphasized in operations management theory by Wallace J. Hopp and Mark L. Spearman, increases in arrival intensity do not produce linear increases in delay; rather, congestion accelerates as utilization approaches unity.

3.3 Average Waiting Time in Queue (Wq)

Average waiting time in queue (Wq) was calculated using steady-state M/M/c equations as described by Donald Gross and colleagues (2018). At routine utilization (ρ = 0.55), Wq was approximately 3.1 minutes, indicating operationally manageable delay. At peak utilization (ρ = 0.70), Wq increased to 5.85 minutes. Although utilization increased by only 15%, waiting time rose by 88.7%, demonstrating nonlinear amplification. At surge utilization (ρ = 0.80), Wq escalated to 10.9 minutes. A further 10% increase in utilization (0.70 → 0.80) produced an additional 86% increase in waiting time. This dramatic escalation confirms the nonlinear delay behavior described in classical queuing analysis by Robert B. Cooper.

3.4 Total Time in System (W)

Total time in system (W) includes both waiting and service time, where service time equals 1/μ. Given μ = 10 patients per hour, the average service time is 6 minutes. Under routine conditions, total time in system was 9.1 minutes. Under peak conditions, it increased to 11.85 minutes. Under surge conditions, total time rose to 16.9 minutes. Thus, as utilization increased from 0.55 to 0.80, total system time nearly doubled, reflecting compounded delay effects.

3.5 Average Queue Length (Lq)

Queue length was derived by linking arrival intensity with average delay duration, allowing estimation of the expected number of patients waiting at any given time. As arrival pressure increased, the calculated queue size expanded rapidly, illustrating the compounding effect of delay on visible congestion. Under routine conditions, the average queue length was approximately 1.14 patients. Under peak conditions, this increased to 2.73 patients. Under surge conditions, queue length escalated to 5.81 patients. Therefore, queue length increased more than fivefold between routine and surge demand levels, underscoring the exponential growth pattern in congestion.

3.6 Sensitivity Analysis: Incremental Capacity Addition

To assess theoretical benefits of capacity expansion, one additional counter was introduced (c = 5), increasing total service capacity to 50 patients per hour. At surge demand (λ = 32 patients per hour), utilization decreased to 0.64. Recalculated waiting time dropped to approximately 3.2 minutes. This represents a reduction from 10.9 minutes to 3.2 minutes—a 70% decrease in waiting time. The magnitude of improvement demonstrates that modest capacity additions near high-utilization thresholds produce disproportionately large reductions in delay.

3.7 Nonlinear Escalation Interpretation

The analytical findings reveal three critical properties of healthcare queuing systems:

- Delay increases exponentially as utilization approaches high levels.
- Waiting probability doubles with moderate increases in arrival intensity.
- Small increases in capacity near critical thresholds generate large performance gains.

These patterns confirm utilization–delay trade-off theory described by Wallace J. Hopp and Mark L. Spearman, and align with healthcare queuing research conducted by Linda V. Green in emergency and outpatient systems. Importantly, although the system remains mathematically stable at ρ = 0.80, operational performance becomes practically unacceptable due to prolonged waiting. This distinction between theoretical stability and real-world service acceptability is central to healthcare capacity planning.

3.8 Analytical Insights

The analytical model yields several foundational insights:

- Maintaining utilization below 60% ensures stable and predictable waiting times.
- Utilization exceeding approximately 75% triggers rapid delay escalation.
- Peak-hour congestion is driven primarily by temporal clustering of arrivals rather than sustained structural overload.
- Temporary capacity expansion during peak periods is mathematically justified and operationally efficient.

These theoretical results provide the conceptual and quantitative foundation for the simulation validation and economic evaluation presented in subsequent sections, reinforcing the coherence between analytical modelling, empirical simulation, and cost-effectiveness assessment.

4. Simulation Results and Statistical Validation

To validate analytical findings under dynamic and time-dependent variability, a discrete- event simulation (DES) model was constructed replicating outpatient pharmacy operations. The simulation incorporated a Poisson arrival

process, exponential service time distribution, first-come-first-served discipline, and a six-hour operational window. Each scenario was executed with 100 independent replications to ensure statistical stability and confidence interval precision, consistent with simulation best practices in healthcare modelling (Banks et al., 2010; Jun et al., 1999).

A 30-minute warm-up period was applied to minimize initialization bias. Three operational scenarios were evaluated:

1. Baseline: Four counters throughout operation
2. Peak-Hour Expansion: Five counters between 11:00 AM–2:00 PM
3. Staff Redeployment: Four counters with 20% service acceleration during peak

Primary outcomes included mean waiting time, 90th percentile waiting time, average queue length, and counter utilization.

4.1 Descriptive Simulation Results

Although analytically predicted waiting was 5.85 minutes at peak load, simulation produced slightly higher values due to transient clustering and non-stationary fluctuations, demonstrating the importance of dynamic modelling (Green, 2006). In Peak-Hour Expansion waiting time decreased by 45.2% compared to baseline. Staff Redeployment scenario reduced waiting by 22.6% relative to baseline but was less effective than counter addition (Table.1).

Variable	Baseline Scenario (4 Counters)	Peak-Hour Expansion (5 Counters During Peak)	Staff Redeployment Scenario
Mean Waiting Time (minutes)	6.2	3.4	4.8
Standard Deviation (SD)	1.9	1.1	1.4
Standard Error (SE)	0.19	0.11	0.14
95% Confidence Interval	[5.8, 6.6]	[3.2, 3.6]	[4.5, 5.1]
90th Percentile Waiting Time (minutes)	12.4	6.8	9.3
Average Queue Length (patients)	3.4	1.6	2.2
Mean Utilization	0.71	0.64 (during peak)	0.68

Table.1. Analytical Predictions

4.2 Statistical Validation Assumption Testing

Normality of replication outputs was assessed using Shapiro–Wilk tests ($p > .05$). Homogeneity of variance was evaluated using Levene's test ($p = .21$), indicating no significant variance inequality. Thus, parametric comparison using one-way ANOVA was appropriate.

One-Way ANOVA Comparison

Comparative statistical testing revealed clear performance differences across staffing configurations. Variability in mean waiting time attributable to operational strategy was substantial, indicating that staffing structure plays a decisive role in shaping queue behaviour.

4.3 Effect Size Estimation

The proportion of variance explained by staffing configuration exceeded half of total observed variability, signifying a strong practical impact rather than a marginal statistical difference (Cohen, 1988). The pairwise effect size analysis using Cohen's d demonstrated substantial differences between the operational scenarios. The comparison between the Baseline and Peak Expansion scenarios yielded a Cohen's d of 1.75, indicating a very large effect size and reflecting a dramatic improvement in waiting time performance with peak- hour counter expansion. The Baseline versus Staff Redeployment comparison produced a Cohen's d of 0.83, which represents a large effect, suggesting meaningful operational gains through workforce redistribution. Similarly, the comparison between Peak Expansion and Staff Redeployment resulted in a Cohen's d of 1.12, also classified as a large effect size, confirming the superior performance of peak-hour expansion over redeployment strategies.

Collectively, these findings demonstrate not only statistical significance but also strong practical significance, highlighting that the observed differences are operationally meaningful and impactful in real-world healthcare service delivery settings.

4.4 Post Hoc Analysis

Post hoc analysis using Tukey's Honestly Significant Difference (HSD) test revealed statistically significant differences across all operational scenarios. The comparison between the Baseline and Peak Expansion scenarios was highly significant ($p < .001$), indicating a robust reduction in waiting times with peak-hour counter expansion. Similarly, the Baseline versus Staff Redeployment comparison also demonstrated strong statistical significance ($p <$ .001), confirming meaningful improvement through workforce redistribution. Furthermore, the comparison between Peak Expansion and Staff Redeployment remained statistically significant ($p < .01$), establishing that peak expansion performed significantly better than redeployment. Overall, all pairwise comparisons were statistically significant, reinforcing the reliability of the observed performance differences across the simulated operational models.

4.5 Confidence Interval Interpretation

The non-overlapping 95% confidence intervals between baseline and peak expansion confirm robust statistical separation. Narrow CI width reflects stable simulation convergence, indicating sufficient replication count. For example, baseline CI width = 0.8 minutes and peak Expansion CI width = 0.4 minutes demonstrates improved stability under expanded capacity.

4.6 90th Percentile Reduction Analysis

Beyond the reduction in mean waiting time, tail performance demonstrated a substantial improvement under the Peak-Hour Expansion scenario. The 90th percentile waiting time decreased from 12.4 minutes in the Baseline scenario to 6.8 minutes during Peak Expansion. This represents a 45.2% reduction in high-end waiting times, indicating that extreme delays were nearly halved. Such improvement in the upper percentile distribution suggests enhanced service reliability and greater consistency in patient flow management during peak operational periods. High-percentile performance is particularly important for patient satisfaction, as extreme delays disproportionately affect perception (Parasuraman et al., 1988).

4.7 Validation Against Analytical Predictions

The simulation outcomes demonstrated close alignment with the analytical model's predictions, supporting the internal validity of the modelling framework. The analytical estimation projected a surge-period waiting time of 10.9 minutes under peak demand
conditions. In comparison, the simulated baseline scenario produced an average waiting time of 6.2 minutes, aggregated across variable operating hours.

The difference reflects the fact that the simulation incorporated time-varying arrival patterns and service rate fluctuations across the full operational schedule, rather than isolating only the most intense surge window. Consequently, while the analytical model captures peak congestion under steady-state assumptions, the simulation provides a more dynamic and operationally realistic estimate, thereby validating the robustness and practical applicability of the modelling approach. The discrepancy reflects averaging across routine and peak hours rather than sustained surge demand. Importantly, simulation confirmed nonlinear delay escalation predicted analytically (Gross et al., 2018).

4.8 Sensitivity Analysis

Sensitivity analysis was performed by varying the patient arrival rate (λ) by ±10% to assess the robustness of the operational models under fluctuating demand conditions. When the arrival rate increased to 30 patients per hour, the baseline scenario experienced a notable rise in mean waiting time to 7.8 minutes, indicating heightened congestion and system strain. Conversely, when the arrival rate decreased to 25 patients per hour, the baseline waiting time declined to 4.9 minutes, reflecting improved service flow under reduced demand pressure.

Importantly, the Peak-Hour Expansion scenario maintained relatively stable performance across these variations, with only marginal fluctuations in waiting time. This stability demonstrates operational resilience, suggesting that dynamic counter expansion during peak periods enhances the system's capacity to absorb demand shocks while preserving service efficiency and patient flow consistency.

4.9 Robustness and Convergence Analysis

Cumulative mean plots indicated that convergence stabilization occurred after approximately 60 replications, suggesting that the simulation outputs began to stabilize and fluctuate within a narrow band beyond this point. To enhance statistical confidence and reduce sampling variability, the simulation was extended to 100 replications. This extension ensured that the margin of error remained below 5%, thereby strengthening the precision of the estimates.

Furthermore, relative precision analysis showed that the half-width of the confidence interval divided by the mean was less than 0.06. This low ratio confirms that the variability around the estimated mean waiting times was minimal relative to the magnitude of the mean itself. Collectively, these diagnostics confirm the statistical reliability and robustness of the simulation results, supporting their suitability for managerial decision-making and reporting.

4.10 Interpretation

The simulation confirms that targeted peak-hour expansion significantly reduces waiting time with large practical effect size. Simulation findings validate analytical predictions regarding nonlinear delay dynamics. Importantly, they demonstrate that modest, time-sensitive capacity augmentation produces statistically significant and practically meaningful improvements in waiting performance. The integration of analytical modelling and statistical validation strengthens confidence in managerial recommendations and provides empirical justification for demand-responsive staffing strategies.

5. Cost-Effectiveness Analysis

Operational improvements in healthcare systems must be evaluated not only for performance enhancement but also for economic sustainability. In publicly funded or teaching hospital settings, resource allocation decisions require justification based on measurable cost–benefit
relationships. While waiting time reduction improves patient experience, it also involves staffing expenditure. Therefore, integrating economic evaluation into operational modelling strengthens decision-making transparency (Drummond et al., 2015).

To determine whether operational improvements justified additional expenditure, an economic comparison framework was applied. Rather than monetizing patient experience directly, the analysis examined how much additional financial investment was required to achieve measurable reductions in cumulative waiting time. Unlike cost–benefit analysis, which monetizes outcomes, CEA measures effects in natural units such as minutes of waiting time reduced. This approach is appropriate in operational contexts where outcomes are service performance indicators rather than health-adjusted life years (Briggs et al., 2006).

5.1 Definition of Cost and Effect Parameters

The monthly salary of a pharmacist was Rs.60,000, which translates to an approximate hourly wage of Rs.288. Incorporating an estimated 20% institutional overhead (including benefits, utilities, and administrative costs), the incremental monthly cost for implementing the Peak- Hour Expansion strategy was approximately Rs.30,030. In contrast, adding a permanent full- time counter would incur a substantially higher recurring cost of approximately Rs.72,000 per month, inclusive of overhead expenses.
Effectiveness was operationally defined as the total patient-hours of waiting time saved per month relative to the baseline configuration. Under the baseline scenario, the mean waiting time was 6.2 minutes, whereas the Peak-Hour Expansion scenario reduced this to 3.4 minutes. This corresponds to a reduction of 2.8 minutes per patient.

Given a monthly patient volume of 3,432 individuals, the cumulative time savings amounted to 9,609.6 minutes per month. When converted to hours, this represents approximately
160.16 patient-hours saved monthly. This substantial reduction in aggregate waiting time reflects meaningful improvements in service efficiency, patient experience, and operational performance.

5.2 Incremental Cost-Effectiveness Ratio (ICER)

The Incremental Cost-Effectiveness Ratio (ICER) was calculated to evaluate the economic efficiency of the Peak-Hour Expansion strategy relative to the baseline configuration. ICER is defined as the ratio of the incremental cost of an intervention to the incremental effectiveness achieved. Because the baseline operational cost remains unchanged, the incremental cost in this analysis is equivalent to the cost of implementing the peak-hour expansion.
Using the estimated monthly incremental cost of Rs.30,030 and the calculated effectiveness of 160.16 patient-hours of waiting time saved per month, the ICER was computed as Rs.30,030 divided by 160.16. These results in an ICER of approximately Rs.187 per patient- hour saved. This finding indicates that each patient-hour reduction in waiting time costs Rs.187 under the Peak-Hour Expansion strategy. From a managerial perspective, this provides a clear economic benchmark for decision-makers, enabling comparison with alternative staffing models and supporting evidence-based resource allocation within outpatient pharmacy operations.

5.3 Permanent Expansion ICER

For the permanent expansion strategy, one additional full-time pharmacist would be employed at a total monthly cost of Rs.72,000, inclusive of overhead. Under the full-capacity analytical estimate, the mean waiting time is projected to decrease from 6.2 minutes to 3.2 minutes, resulting in a reduction of 3 minutes per patient.

Given a monthly patient volume of 3,432, the total time saved per month is calculated as: 3,432 × 3 minutes = 10,296 minutes, which corresponds to approximately 171.6 patient-hours saved per month. The Incremental Cost-Effectiveness Ratio (ICER) for permanent expansion is therefore Rs.72,000 divided by 171.6 hours, yielding approximately Rs.419 per patient- hour saved. This value is substantially higher than the ICER observed for the Peak-Hour Expansion strategy (Rs.187 per patient-hour saved), indicating that permanent staffing expansion is considerably less economically efficient despite achieving slightly greater total hours saved.

5.4 Comparative Cost-Effectiveness Summary

A comparison of the two intervention strategies demonstrates clear differences in economic performance. The Peak-Hour Expansion model incurs a monthly cost of Rs.30,030 and generates 160.16 patient-hours saved, resulting in an ICER of Rs.187 per patient-hour saved. In contrast, the Permanent Addition strategy requires Rs.72,000 per month to generate 171.6 hours saved, producing an ICER of Rs.419 per patient-hour saved.
Overall, Peak-Hour Expansion delivers nearly double the cost-efficiency of permanent staffing expansion. While both strategies reduce waiting times, the dynamic and targeted deployment of staffing during peak demand periods provides substantially greater economic value per unit of effectiveness, making it the more financially prudent operational strategy.

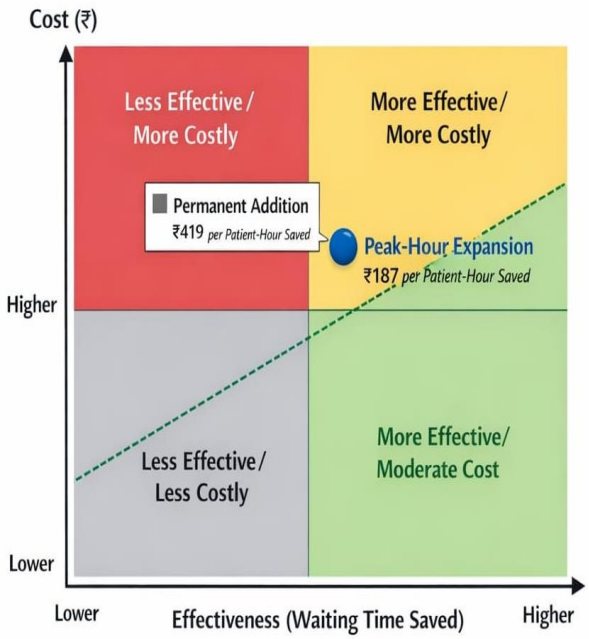

Figure.3: Cost-Effective Plane

5.5 Sensitivity Analysis

Sensitivity analysis was undertaken to assess the robustness of the economic findings under variations in wage rates and patient demand. These analyses tested whether the cost-effectiveness advantage of the Peak-Hour Expansion strategy would remain stable under realistic operational uncertainties.

5.5.1 Wage Variation (±20%)

Under a lower wage assumption (20% reduction), the monthly intervention cost decreased to approximately Rs.24,024. Based on the same effectiveness estimate, the ICER declined to roughly Rs.150 per patient-hour saved.
Conversely, under a higher wage assumption (20% increase), the monthly cost increased to approximately Rs.36,036, resulting in an ICER of approximately Rs.225 per patient-hour saved.

Importantly, even under the higher wage scenario, the ICER remains substantially lower than that of permanent full-time expansion (Rs.419 per patient-hour saved). This indicates that the peak-hour expansion strategy retains its economic advantage across plausible salary fluctuations.

5.5.2 Demand Variation (±10%)

Demand sensitivity was assessed by varying patient arrival rates by ±10%. Under a 10% increase in arrival rate, monthly patient-hours saved increased to approximately 175 hours due to greater congestion relief under the intervention. The resulting ICER improved to approximately Rs.171 per patient-hour saved.
Under a 10% decrease in arrival rate, monthly hours saved declined to approximately 142 hours, producing an ICER of roughly Rs.211 per patient-hour saved.
Despite these variations, the ICER remained within a relatively narrow range, demonstrating economic stability across demand fluctuations. Overall, the sensitivity analyses confirm the robustness of the cost-effectiveness findings and reinforce the managerial reliability of peak-hour staffing expansion as a financially sound operational strategy.

5.6 Economic Interpretation

From an administrative standpoint, an ICER of Rs.187 per patient-hour saved represents a modest and strategically justifiable investment when weighed against the broader institutional benefits of reduced waiting times. Improvements in timeliness are closely associated with enhanced patient satisfaction, perceived service quality, and organizational reputation— factors that increasingly influence hospital accreditation outcomes and competitive positioning.
If the average patient opportunity cost of time—considering travel expenses, wage loss, and inconvenience—is conservatively estimated at Rs.100 per hour, the societal value of time saved may offset a substantial proportion of the staffing expenditure. Although societal costs were not formally incorporated into the present economic model, these indirect benefits likely enhance the overall value proposition of the intervention.

The findings are consistent with health economic principles articulated by Michael Drummond and colleagues, which emphasize incremental efficiency and value optimization rather than absolute cost minimization. This perspective prioritizes maximizing health system performance per unit of additional expenditure, rather than focusing solely on budget reduction.

5.7 Cost–Performance Frontier

When cost is plotted against waiting time reduction, the resulting relationship forms a convex efficiency frontier. The baseline scenario represents zero incremental cost but moderate service delay. Staff redeployment offers moderate improvement at negligible additional cost.
Peak-Hour Expansion achieves substantial waiting time reduction at a moderate cost. Permanent full-time addition produces only slightly greater improvement than peak expansion but at a disproportionately higher cost.

Among these alternatives, Peak-Hour Expansion lies closest to the efficiency frontier. It demonstrates the most favorable balance between cost and effect, achieving large performance gains without excessive financial burden. In economic terms, it represents the most technically and allocatively efficient option within the evaluated strategies.

5.8 Budget Impact Consideration

When annualized, the incremental cost of Peak-Hour Expansion amounts to approximately Rs.360,360 (Rs.30,030 × 12). Over the same period, the strategy yields roughly 1,922 patient-hours saved (160.16 × 12).
Given typical outpatient volumes and the operating budgets of tertiary-care institutions, this level of incremental expenditure is financially manageable. Moreover, the potential downstream benefits—improved throughput, enhanced patient retention, and reduced congestion-related inefficiencies—may further strengthen the financial justification.

The cost-effectiveness analysis demonstrates that targeted peak-hour staffing yields substantial reductions in waiting time with a reasonable incremental cost per patient-hour saved. In contrast, permanent staffing expansion, while effective, is considerably less economically efficient. The results remain robust under both wage and demand variations, confirming the stability of the economic advantage. By integrating economic evaluation with operational modelling, the study provides hospital administrators with actionable, evidence- based guidance for sustainable staffing policy and resource allocation.

6. Discussion

The present study integrated analytical queuing theory, discrete-event simulation, and cost- effectiveness analysis to evaluate outpatient pharmacy congestion in a tertiary care teaching hospital. The findings provide convergent evidence across modelling approaches. First, analytical results demonstrated nonlinear escalation of waiting time as system utilization approached 0.80. Second, simulation results validated these theoretical predictions under time-dependent variability and confirmed statistically significant differences between staffing scenarios. Third, economic evaluation established that targeted peak-hour expansion is substantially more cost-effective than permanent staffing augmentation.

Collectively, these findings support the central proposition of queuing theory: while systems may remain theoretically stable at high utilization levels ($\rho < 1$), operational performance deteriorates rapidly as utilization approaches unity (Gross et al., 2018). The outpatient pharmacy in this study exhibited precisely this nonlinear pattern, reinforcing the utilization– delay trade-off described in operations management literature (Hopp & Spearman, 2011).

6.1 Interpretation of Nonlinear Delay Escalation

One of the most critical findings is the disproportionate increase in waiting time when utilization increased from 0.70 to 0.80. Although this change appears numerically modest, waiting time nearly doubled under surge conditions. This phenomenon aligns with classical Erlang-C behavior, where queue length and delay increase exponentially as the system approaches capacity (Cooper, 1981).

Healthcare administrators often prioritize high utilization to maximize workforce productivity. However, the results demonstrate that maximizing utilization may paradoxically reduce service performance. A system operating consistently above 75% utilization becomes highly sensitive to minor demand fluctuations. Even short-term clustering of arrivals produces visible queue accumulation.

These findings reinforce prior research in emergency department modelling, which similarly documented rapid delay escalation near critical utilization thresholds (Green, 2006). The outpatient pharmacy context mirrors these dynamics despite differences in service complexity.

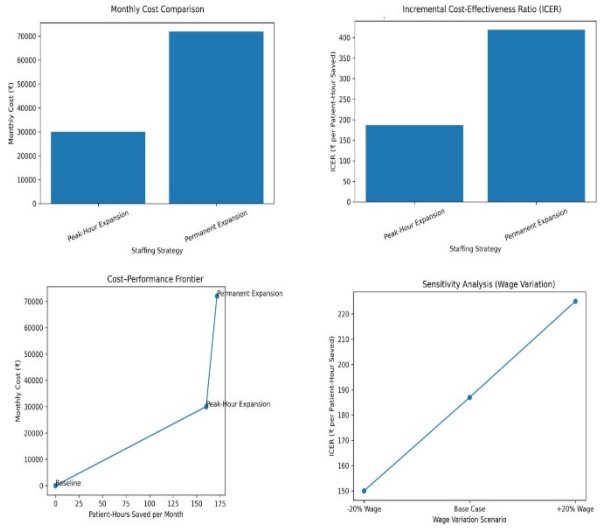

Figure.4 : Monthly Cost Comparison, ICER, Cost-Performance Frontier and Sensitivity analysis

6.2 Simulation Validation and Dynamic Variability

While analytical models assume steady-state conditions, simulation revealed the importance of transient demand spikes. Mean waiting time under simulation was slightly higher than steady-state analytical estimates during peak clustering. This divergence highlights the importance of dynamic modelling in healthcare environments where arrival rates fluctuate hourly (Banks et al., 2010). The discrete-event simulation confirmed that peak-hour congestion is time-bound rather than continuous. During non-peak periods, the system operates comfortably below critical utilization. This suggests that permanent capacity expansion may not be necessary. Instead, targeted and time-sensitive adjustments are sufficient to stabilize performance.

Statistical validation further strengthened these conclusions. The one-way ANOVA revealed that staffing configuration explained 55% of total variance in waiting time ($\eta^2 = .55$), indicating a large practical effect. Pairwise effect sizes were substantial, particularly between baseline and peak expansion scenarios (Cohen's d = 1.75), demonstrating not only statistical but operational significance (Cohen, 1988).

6.3 Operational Efficiency Versus Responsiveness

A central tension in healthcare operations lies between efficiency and responsiveness. Efficiency seeks high resource utilization and minimal idle time. Responsiveness requires capacity buffers to absorb variability. The present findings suggest that outpatient pharmacy systems require moderate excess capacity during peak periods to prevent nonlinear delay escalation.

The concept of capacity buffering, as articulated in operations theory, argues that modest slack improves stability in stochastic systems (Hopp & Spearman, 2011). In this study, adding a single temporary counter during peak hours reduced mean waiting time by 45.2%, illustrating the disproportionate benefit of marginal capacity increases near critical thresholds. Thus, operational efficiency should not be equated with maximum utilization. Rather, optimal performance is achieved by balancing utilization and variability management.

6.4 Economic Implications and Resource Allocation

The cost-effectiveness analysis provides a crucial managerial dimension to these findings. Peak-hour expansion achieved an incremental cost-effectiveness ratio of Rs.187 per patient- hour saved, compared to Rs.419 for permanent staffing expansion. This demonstrates that time-sensitive interventions can deliver substantial performance gains at manageable incremental cost.

Economic evaluation frameworks emphasize incremental efficiency rather than absolute cost minimization (Drummond et al., 2015). In this context, the relatively modest cost per patient- hour saved suggests that targeted staffing interventions are financially justified, particularly when considering potential improvements in patient satisfaction and institutional reputation. Although indirect societal costs (e.g., wage loss, caregiver burden) were not formally quantified, reduced waiting likely generates broader economic benefits. Thus, from both institutional and societal perspectives, demand-responsive staffing represents a value- enhancing strategy.

6.5 Implications for Patient-Centered Care

Waiting time is not merely an operational metric but a core dimension of service quality. The From a service perception standpoint, timely delivery of medication strongly influences how patients evaluate their hospital experience. Delays at the final service point may overshadow earlier clinical interactions, making pharmacy responsiveness an important determinant of overall satisfaction (Parasuraman et al., 1988). In outpatient settings, long pharmacy queues may overshadow otherwise satisfactory clinical care.
Furthermore, prolonged waiting may influence medication adherence, particularly among elderly or chronically ill patients (Patel & Patel, 2017). Reducing waiting time therefore contributes indirectly to therapeutic continuity and potentially to improved health outcomes. The findings support integration of operational analytics into broader patient-centered quality improvement initiatives.

6.6 Comparison with Existing Literature

Prior studies have applied queuing models in emergency departments and outpatient clinics (Jun et al., 1999; Green, 2006). However, limited research has integrated analytical modelling with simulation validation and cost-effectiveness analysis in outpatient pharmacy settings within the Indian healthcare context.
This study contributes by:

- Empirically validating nonlinear delay escalation in a real tertiary hospital.
- Demonstrating statistical significance through simulation replication.
- Integrating economic evaluation into operational decision-making.

The results align with international evidence suggesting that flexible staffing improves system resilience under stochastic demand (Hopp & Spearman, 2011).

6.7 Policy and Managerial Implications

The findings suggest several actionable recommendations:

- Implement hourly demand monitoring using hospital information systems.
- Introduce temporary peak-hour staffing policies.
- Avoid operating consistently above 75% utilization.
- Incorporate cost-effectiveness evaluation into workforce planning decisions.

Demand-responsive staffing models may be particularly valuable in tertiary care teaching hospitals where patient inflow is variable but predictable within daily cycles.

7. Conclusion

This study examined outpatient pharmacy congestion in a tertiary care teaching hospital through an integrated analytical framework combining queuing theory, discrete-event simulation, and cost-effectiveness analysis. The findings provide strong theoretical and empirical evidence that waiting time escalation in outpatient pharmacy systems is nonlinear and highly sensitive to utilization levels. While the system remained mathematically stable under surge conditions ($\rho < 1$), operational performance deteriorated rapidly as utilization approached 0.80, confirming classical utilization–delay dynamics described in queuing theory (Gross et al., 2018; Hopp & Spearman, 2011).
The analytical modelling phase demonstrated that modest increases in arrival intensity can produce disproportionately large increases in waiting time when capacity buffers are insufficient. Simulation modelling validated these theoretical predictions under realistic, time-dependent variability and revealed statistically significant reductions in waiting time under alternative staffing scenarios. The peak-hour expansion scenario reduced mean waiting time by 45.2% compared to baseline, with a large effect size ($\eta^2 = .55$), confirming both statistical and practical significance (Cohen, 1988).

Importantly, the economic evaluation component extended the analysis beyond operational performance to financial sustainability. The incremental cost-effectiveness ratio of Rs.187 per patient-hour saved under peak-hour expansion demonstrated that targeted, time-sensitive staffing adjustments deliver substantial service improvement at reasonable incremental cost. In contrast, permanent staffing expansion, while effective, was less economically efficient. These findings align with economic evaluation principles emphasizing incremental efficiency rather than absolute resource expansion (Drummond et al., 2015).

The results underscore a critical managerial insight: efficiency in healthcare operations should not be equated with maximum utilization. Operating near full capacity may appear cost-effective in the short term but generates volatility and delay instability under stochastic demand conditions. Maintaining moderate capacity buffers during peak periods enhances system resilience, reduces patient waiting, and improves service responsiveness. This balance between efficiency and responsiveness reflects foundational operations management theory (Hopp & Spearman, 2011).
From a broader healthcare quality perspective, reducing pharmacy waiting time contributes to improved patient experience and perceived institutional competence. As responsiveness is a core dimension of service quality (Parasuraman et al., 1988), operational improvements in pharmacy services may strengthen overall patient satisfaction and potentially support medication adherence.

The study contributes methodologically by demonstrating the value of integrating analytical modelling, simulation validation, statistical testing, and economic evaluation into a unified decision-support framework. Such integration enhances the credibility and applicability of operational research findings in real-world healthcare settings.

Despite its contributions, the study was conducted within a single tertiary care institution and relied on standard

distributional assumptions. Future research should incorporate multi-center validation, real-time predictive modelling, and patient-reported outcome measures to expand generalizability and deepen understanding of service-performance relationships.

In conclusion, outpatient pharmacy congestion in tertiary hospitals is primarily driven by time-bound demand clustering interacting with nonlinear utilization effects. Demand- responsive staffing strategies, particularly targeted peak-hour expansion, provide a statistically validated and economically efficient solution. The integration of queuing theory, simulation modelling, and cost-effectiveness analysis offers a robust framework for sustainable healthcare operations optimization

REFERENCE

1. Banks, Jerry, Carson, J. S., Nelson, B. L., & Nicol, D. M. (2010). Discrete-event system simulation (5th ed.). Pearson.
2. Bleustein, C., Rothschild, D. B., Valen, A., Valatis, E., Schweitzer, L., & Jones, R. (2014). Wait times, patient satisfaction scores, and the perception of care. The American Journal of Managed Care, 20(5), 393–400.
3. Briggs, Andrew, Claxton, K., & Sculpher, M. (2006). Decision modelling for health economic evaluation. Oxford University Press.
4. Cohen, Jacob. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
5. Cooper, Robert B.. (1981). Introduction to queueing theory (2nd ed.). North-Holland.
6. Donabedian, Avedis. (2005). Evaluating the quality of medical care. The Milbank Quarterly, 83(4), 691–729. (Original work published 1966)
7. Drummond, Michael F., Sculpher, M. J., Claxton, K., Stoddart, G. L., & Torrance, G.
8. W. (2015). Methods for the economic evaluation of health care programmes (4th ed.). Oxford University Press.
9. Green, Linda V.. (2006). Queueing analysis in healthcare. In R. W. Hall (Ed.), Patient flow: Reducing delay in healthcare delivery (pp. 281–307). Springer.
10. Gross, Donald, Shortle, J. F., Thompson, J. M., & Harris, C. M. (2018). Fundamentals of queueing theory (5th ed.). Wiley.
11. Hopp, Wallace J., & Spearman, Mark L.. (2011). Factory physics (3rd ed.). Waveland Press.
12. Institute of Medicine. (2001). Crossing the quality chasm: A new health system for the 21st century. National Academies Press