International Journal of Basic and Applied Science 13 (4) (2025) 179-156

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Queueing theory and simulation for reducing patient waiting time in emergency departments

Aminah [1], Harauly Lady Lusiana Manalu [2], Verawaty Fitrinelda Silaban [3], Dewi Sartika Munthe [4], and Rahmaini Fitri Harahap [5]

[1,2,3,4,5] Keperawatan Dan Kebidanan, Universitas Prima Indonesia, Indonesia

Article Info

Article history:

Received Oct 19, 2024
Revised Dec 20, 2024
Accepted Feb 22, 2025

Keywords:

Accumulative priority queue;
APQ-h;
Discrete event simulation;
Emergency Departments (EDs);
Patient waiting time;
Service flow management.

ABSTRACT

Emergency Departments (EDs) are increasingly overwhelmed by rising patient volumes and limited service capacity, leading to long waiting times and reduced care quality. This study addresses the inefficiencies of conventional queue policies by proposing a dynamic scheduling approach known as the Accumulative Priority Queue with Finite Horizon (APQ-h). APQ-h integrates time-based priority accumulation with triage thresholds, allowing for a more realistic representation of how clinicians manage patient flow. Using discrete-event simulation and simulation-based optimization, the research calibrates accumulation rate parameters (β) to minimize total waiting time (TWT) and ensure compliance with clinical response time targets (APT). A real-world case study and sensitivity analysis reveal that optimal configurations of β enable balanced and adaptive queue management without disadvantaging any patient group. The findings contribute a hybrid queue discipline that bridges mathematical modeling and clinical practice, offering practical implications for improving ED throughput and resource utilization. Although the simulation relies on stylized assumptions, it opens avenues for real-time validation, integration with adaptive triage systems, and scalability across diverse healthcare settings.

Corresponding Author:

Aminah,
Keperawatan Dan Kebidanan,
Universitas Prima Indonesia,
Jl. Sampul No 3 Sei Putih Barat Kec. Medan Petisah, Kota Medan, Sumatera Utara, 20118, Indonesia
Email: aminah@unprimdn.ac.id

1. INTRODUCTION

Emergency Departments (EDs) are one of the most vital units in the health care system, especially in countries with an aging population and a high burden of chronic diseases[1]. In recent decades, there has been a significant increase in patient visits to the Emergency Department, while service capacity has not experienced comparable growth[2], [3]. This imbalance between demand and capacity results in overcrowding, which has an impact on increasing patient waiting times, decreasing service quality, and increasing risk of adverse events[4].

To answer these challenges, an effective and adaptive patient flow management strategy is needed[5]. One of the approaches that is widely studied in the field of operations research is queueing theory, which provides a mathematical framework for modeling service systems that are stochastic and complex such as Emergency Departments[6], [7]. When combined with discrete-event simulation techniques, this approach allows for the evaluation of system performance in a variety of realistic operational scenarios[8].

Traditional queuing policies in emergency departments generally use a pure priority rules, where patients with higher severity always come first[9], [10]. Although simple and easy to implement, this approach often does not reflect real practice in the field, especially in crowded Emergency Department conditions[11]. Empirical studies show that decision-makers in the Emergency Department consider not only the severity, but also the actual waiting time of the patient[10], [12]. This behavior is in accordance with the concept of accumulative priority queue (APQ), where patients accumulate priority points as the waiting time increases[9], [13].

In this study, an extension of the APQ model, namely Accumulative Priority Queue with Finite Horizon (APQ-h), which includes an upper limit of waiting time for priority accumulation, is examined. After crossing this limit, patients no longer accumulate additional points, so this approach is closer to the patient management policy that is applied empirically in various emergency departments.

This study aims to analyze the effectiveness of APQ-h in reducing patient waiting time in the Emergency Department through a simulation-based optimization approach. Through a case study in a real hospital and an analysis that is extended to various simulation scenarios, this article shows how the integration between queue theory and simulation can be used as a tool in strategic decision-making to improve service performance in the Emergency Departments. This paper evaluates the effectiveness of APQ-h through a simulation-based optimization framework, aiming to determine optimal accumulation rates (β) that reduce patient waiting times while maintaining clinical compliance.

2. RESEARCH METHOD

The problem of patient waiting time in the Emergency Department (EDs) has become a major concern in various operational studies and health service management[14], [15]. Various operations research-based approaches, particularly queueing theory and discrete-event simulation, have been developed to model, analyze, and optimize patient flows in complex and dynamic Emergency Departments systems[16].

In general, queuing policies in the Emergency Department can be grouped into two main approaches: first-in–first-out (FIFO) and priority based on patient severity[17], [18]. In practice, a priority-based approach is more commonly used, where patients with more severe conditions will take precedence over patients with mild conditions. This policy is also widely used in theoretical and simulation models developed in previous literature [19], [20], [21].

However, some studies have shown that *the pure priority* approach has its drawbacks, especially in situations with high patient arrival rates. Patients with low priority can experience very long, even neglected delays in the system. To overcome these weaknesses, a dynamic prioritization approach was developed, one of which is the Accumulative Priority Queue (APQ), as introduced by Stanford et al. (2014)[22]. In APQ, patients accumulate *priority points* during their waiting, with the rate of accumulation determined by their severity[23]. This allows patients with lower priority, but with longer wait times, to move up the queue order[24].

Other studies have also developed dispatching rule-based approaches, such as *the* $c\mu$ rule known in manufacturing contexts and have been adapted in healthcare settings [25]. In this rule, priority is given based on a combination of severity and efficiency of service, assuming that waiting costs increase linearly or convection with time. However, this approach often relies on asymptotic assumptions or heavy *traffic conditions*, so its application in the real context of the Emergency Departments is limited.

An empirical study conducted by Ding et al. (2019) shows that in practice, decision-makers in the Emergency Departments consider target time limits based on triage systems, such as the Canadian Triage Acuity Scale (CTAS), as well as the actual wait time of patients[10][26]. These findings inspired the development of a new policy, the Accumulative Priority Queue with Finite Horizon (APQ-h), which combines APQ flexibility with priority accumulation limits based on triage time targets[22][9]. In this policy, the patient only collects priority points for a certain time, after which the priority value does not increase again.

A simulation-based approach based on discrete events (DES) has been widely used to evaluate the performance of queue policies in the Emergency Departments[27], [28]. The simulation model allows for a more realistic representation of the variability of patient arrival, service time, additional testing needs, as well as the patient's re-evaluation process. Ferrand et al. (2016), for example, show that dynamic queue strategies such as APQ are superior to traditional *fast track* approaches, especially in high-density conditions.

Overall, previous literature confirms that queue theory and simulation have great potential in aiding managerial decision-making in Emergency Departments. However, there is a need to develop models that are more in line with clinical practice, take into account resource limitations, and maintain a balance between operational efficiency and quality of service. This research is here to answer this need by evaluating the APQ-h policy through simulations that represent the real conditions of the Emergency Departments and comparing it with conventional queue policies.

2.1 Patient Routing in Emergency Departments

Patient routing management in the Emergency Departments is a fundamental component in the design of an efficient queue system[7], [15], [20], [29]. The service process in the Emergency Departments is dynamic and stochastic, characterized by variability in the time of patient arrival, the severity of medical conditions, and different treatment paths. Therefore, an accurate representation of the patient's flow is essential for building accurate queue and simulation models.

In general, patients who come to the Emergency Departments first go through an administrative registration process and then undergo a triage process, which classifies patients into several priority levels based on the severity of their condition. In the context of this study, the triage system is assumed to group patients into three priority levels, namely[29]:

a) High Priority (HP)
b) Medium Priority (MP)
c) Low Priority (LP)

After triage, the patient enters the first service stage in the form of an initial consultation with a doctor (first consultation). The results of this consultation can be in the form of:

a) The patient is immediately discharged home or referred to hospital, or
b) Patients need follow-up examinations (such as laboratory, radiology, or specialist consultation) before re-entering the queue for a second consultation.

Thus, there are two main stages of medical services that need to be considered in the model:

a) 1C (First Consultation): First consultation after triage.
b) 2C (Second Consultation): Re-consultation after additional checks.

To form a queue system that reflects this complexity, patients are grouped into six queue categories, namely:

a) 1C-HP, 1C-MP, 1C-LP: Patients first consultation by priority.
b) 2C-HP, 2C-MP, 2C-LP: Second consultation patients based on priority.

When a doctor completes one consultation, the system must determine the next patient to be served from all six categories. Therefore, queue discipline strategies have a crucial role in determining system performance.

Mathematically, this queue system can be modeled as a multi-class, multi-phase priority queue with feedback. Suppose:

a) λ_i is the rate of patient arrival with priority i, with $i \in \{\text{HP}, \text{MP}, \text{LP}\}$
b) α_i is the probability that the priority *i patient* will be discharged after the first consultation.
c) $1 - \alpha_i$ is the probability that the patient will need a second consultation.
d) τ_i is the wait time for priority i patients for the first consultation.
e) ν_i is the wait time for priority I patients for the second consultation.

Thus, the Total Waiting Time (TWT) for priority i patients is expressed as[30][15][20]:

$$\mathbb{E}[TWT_i] = \alpha_i \cdot \mathbb{E}[\tau_i] + (1 - \alpha_i) \cdot \mathbb{E}[\tau_i - \nu_i] \quad (1)$$

The average waiting time for the first and second consultations depends on the queue policy implemented and the overall system load. The model also includes a feedback component, which is the probability of a patient re-entering the system after a medical test.

The patient process flow diagram visually depicts this structure, with the connecting arrow from triage to queue 1C, proceeding to possible discharge or entry into queue 2C, and finally exiting the system.

The model also considers that service capacity (number of doctors and examination rooms) is limited, so the selection of the next patient when a doctor is available will have a direct impact on the system's key performance indicators (KPIs), such as:

a) Arrival to Provider Time (APT): time from the patient to the Emergency Departments until the doctor treats him.
b) Total Waiting Time (TWT): total waiting time from the first consultation to discharge from the Emergency Departments.
c) Length of Stay (LoS): total time of the patient's presence in the Emergency Departments.

Taking into account different service lines and priority levels, this representation allows for more realistic testing and optimization of queue management strategies through a simulation approach.

2.2 Key Performance Indicators

Measuring the performance of the service system in the Emergency Departments requires indicators that are valid, relevant, and able to reflect the efficiency and quality of the health services provided[31]. In the context of queue management and patient flow simulation, key performance indicators (KPIs) serve as the basis for evaluating the various patient flow management policies implemented.

Some of the KPIs that are widely used in the literature and clinical practice include[32]:

a) Arrival to Provider Time (APT)

APT or also known as "door-to-doc" is the time between the patient's arrival at the Emergency Departments and the first time it is treated by medical personnel (doctors). This indicator is especially important because many medical conditions are time-sensitive, such as myocardial infarction or stroke, that require immediate intervention.

Each patient priority level has a maximum time limit (T_i) established based on a triage system, e.g. the Canadian Triage Acuity Scale (CTAS):

$$APT_i = t_{\text{provider}} - t_{\text{arrival}}, \text{ for priority patients } i \quad (2)$$

Performance targets are also set in the form of the maximum percentage of patients allowed to exceed the time limit, for example:

$$\mathbb{P}\ (APT_i > T_i \leq P_i \quad (3)$$

where P_i is the maximum proportion allowed according to the performance standards.

b) Total Waiting Time (TWT)

Total Waiting Time includes the accumulated waiting time of the patient from arrival to completion of undergoing the entire medical consultation process. Patients can go through one or two consultations, depending on the complexity of the case and diagnostic needs.

As described in Subsection 2.1, the expected value of the TWT for patients with priority i is:

$$\mathbb{E}[TWT_i] = \alpha_i \cdot \mathbb{E}[\tau_i] + (1 - \alpha_i) \cdot \mathbb{E}[\tau_i + \nu_i] \quad (4)$$

where:

α_i is the probability of the patient being discharged after the first consultation,

τ_i and ν_i The lead time for the first and second consultations, respectively.

This indicator greatly affects the perception of service quality felt by patients.

c) Length of Stay (LoS)

LoS is the total duration from the time the patient comes to the Emergency Departments until discharge (discharge or referral). LoS reflects the efficiency of the system in completing the entire patient care process. Although closely related to TWT, LoS also includes time spent on additional examinations, waiting for laboratory results, as well as clinical observations.

$$LoS = TWT + T_{service} + T_{diagnostic} \quad (5)$$

Increasing LoS can be an indicator of system density, limited resource capacity, or inefficiency in the service process.

d) Congestion Level
Although not directly included in formal KPIs, density levels—measured through average occupancy rates (ρ) and queue lengths—can be used to measure the risk of system overload. Overcrowding is associated with an increased risk of medical errors, labor stress, and decreased patient safety (Weiss et al., 2004).

e) Proportion of Patients Exceeding APT Limit
In the context of an evaluation of queue management policies, the proportion of patients with APT who exceeded the time limit T_i It is also used as a critical indicator. For example, for priority 4 patients with a time limit of 60 minutes and a maximum target of 15%, then:

$$\Delta_i = \max\{\mathbb{P}(APT_i > T_i) - P_i, 0\} \quad (6)$$

Value Δ_i used in the formulation of objective functions in optimization to measure how far actual performance deviates from the desired target.

2.3 Patient Flow Management Policies

A patient flow management policy (queue discipline or dispatching rule) is a set of rules that determine the order of patient services by medical personnel in the Emergency Departments when there is more than one patient waiting. In a complex queue system such as the Emergency Departments, the selection of the right flow management policy greatly determines the performance of the system, both in terms of operational efficiency and the quality of patient service[33][29][15].

As previously explained, the queue in the Emergency Departments consists of several categories of patients, classified by severity (high, medium, low priority) and service stage (first and second consultations). Thus, the queue system consists of six categories of patients:

a) 1C-HP, 1C-MP, 1C-LP: High, medium, low priority patients waiting for the first consultation.
b) 2C-HP, 2C-MP, 2C-LP: High, medium, low priority patients waiting for a second consultation.

Each time a doctor completes a consultation, the system must determine one of the six categories as the next patient. This selection is carried out based on the queue management policy implemented. The two major policy groups that will be described are:

Pure Priority Rules

A pure priority policy is a conventional approach in which patients are served on a fixed priority basis, without considering actual waiting times. In this policy, patients from the highest priority category who queue first will be served first[10][34].

Some variations of pure priority rules that are commonly used include:

a) PR-1C (First Consultation Priority Rule): Give absolute priority to all patients waiting for the first consultation. Order of service:
1C-HP → 1C-MP → 1C-LP → 2C-HP → 2C-MP → 2C-LP.
b) PR-2C (Second Consultation Priority Rule): Giving priority to patients waiting for a second consultation so that they can leave the system immediately. Order:
2C-HP → 2C-MP → 2C-LP → 1C-HP → 1C-MP → 1C-LP.
c) PR-AI (Acuity Index Rule): Prioritize the severity of the patient, and in each priority, the first consultation comes first. Order:
1C-HP → 2C-HP → 1C-MP → 2C-MP → 1C-LP → 2C-LP.
d) PR-HN (Hybrid Rule - Hospital Norm): The rule, which is applied by the majority of medical personnel in hospital case studies, combines PR-AI for high-priority patients and PR-2C for medium- and low-priority patients.

Each of these rules reflects a different orientation to system performance:

a) PR-1C focus on APT decline (arrival to provider time),
b) PR-2C focus on reducing total wait times and rapid patient discharge,
c) PR-AI Balancing the quality of service based on severity,

d) PR-HN Represent commonly found empirical practices.

Accumulative Priority Queues (APQ and APQ-h)

To overcome the rigidity of the pure priority policy, a more flexible and dynamic Accumulative Priority Queue (APQ) approach was developed. In APQ, patients accumulate priority points during the wait, with different accumulation rates for each category of patients. The patient with the highest points will be selected for further treatment[9], [22], [35].

Mathematically, if the category i patient arrives at the time of t_0, and the rate of accumulation of points is β_i, then at time t, the priority points are:

$$PP_i(t) = \beta_i \cdot (t - t_0) \tag{7}$$

The next selection of patients is carried out by:

$$Select\ patients\ with\ \max\{PP_i(t)\} \tag{8}$$

The advantage of APQ is the ability to balance between the severity of the patient and the length of the waiting time. For example, a low-priority patient who has been waiting for a long time can score higher points than a newly arrived medium-priority patient.

APQ-h: Accumulative Priority Queue with Finite Horizon

To better reflect real practices, a variant of APQ with Finite Horizon (APQ-h) was developed. In this approach, priority points are no longer accumulated after the waiting time reaches the maximum limit determined by the triage system. In other words, the patient only collects points until the time T_i, and after that the value remains.

$$PP_i(t) = \begin{cases} \beta_i \cdot (t - t_0), & if\ \ t - t_0 \le T_i \\ \beta_i \cdot T_i, & if\ \ t - t_0 \le T_i \end{cases} \tag{9}$$

by limiting the accumulation of points, the APQ-h approach reflects clinical behaviors in which after exceeding a safe time limit, the patient is deemed to need to be treated immediately without further special treatment.

2.4 Determination of the Optimal APQ-h Discipline

Determining optimal queue discipline in the Accumulative Priority Queue with Finite Horizon (APQ-h) approach is a crucial step in the implementation of an effective patient flow management strategy in the Emergency Department[9], [22]. APQ-h allows for dynamic adjustment of the order of care based on a combination of patient severity and wait time, but with an upper limit on the accumulation of priority points determined based on triage policies or clinical guidelines.

To achieve optimal system performance, it is necessary to calibrate the parameter β, which is the rate of accumulation of priority points for each patient category (based on priority and consultation stage). The optimal value of parameter β will balance between two main objectives: Adherence to the initial service time limit (APT) according to clinical standards and Minimization of total patient waiting time (TWT) in the system.

a) Optimization Goal Formulation

As previously explained, the objective function used is multi-criteria and stochastic, as it considers more than one key performance indicator that is random. Formally, objective functions are expressed as:

$$\min_{\beta} \left\{ W \cdot \sum_{i=1}^{3} u_i \lambda_i \cdot \Delta_i + \sum_{i=1}^{3} v_i \lambda_i \cdot \mathbb{E} \cdot [TWT_i] \right\} \tag{10}$$

With:

$\Delta_i = max\{\mathbb{P}(APT_i > T_i) - P_i, 0\}$

$\mathbb{E} \cdot [TWY_i]$: Expected Lead Time for Total Priority Patients *i;*

u_i, v_i The Weight of Each Criterion;

λ_i : Priority Patient Arrival Rate *i;*

W: weighting coefficient between APT penalty and total waiting time.

This objective function combines adherence to clinical time targets and operational efficiency, thus reflecting the complexity of decision-making in a real Emergency Department environment.

b) Setting APQ-h Parameters

In the APQ-h system, priority points for each patient are calculated as:

$$PP_i(t) = \begin{cases} \beta_i \cdot (t - t_0), & if \ \ t - t_0 \le T_i \\ \beta_i \cdot T_i, & if \ \ t - t_0 \le T_i \end{cases} \quad (11)$$

With:

β_i : Rate of accumulation for patient category *i;*

T_i: Maximum Time Limit of Accumulation Based on Triage System;

t_0: Patient Arrival Time;

t : the time when the service decision is made;

Value β_i The optimal should be determined in such a way as to allow patients with lower priority but long wait times to still have a chance to be served, without sacrificing newly arrived high-priority patients.

c) Solution Search Strategy

Since the explicit form of objective functions is unknown and performance evaluation can only be done through simulation, the optimization method used is Simulation-Based Optimization (SBO). The SBO process includes:

1) Definition of solution space: Parameters β are collectively restricted to maintain balance (e.g. $\sum \beta_i = 10$).
2) Simulation-based evaluation: Any configuration β tested in simulations to derive values from relevant KPIs.
3) Adaptive iteration: Optimization algorithms generate new configurations based on previous performance (e.g. through scatter search or genetic algorithms approaches).

d) Clinical Success and Implementation Criteria.

The optimal solution of β parameters is considered successful if:

1) The proportion of patients who exceed the APT deadline for each priority does not exceed the tolerance threshold value P_i.
2) The average TWT for all patient categories was at a minimal level.
3) The system is able to adapt to changes in density without causing an imbalance of services between priorities.

In addition to the quantitative aspect, the resulting configuration must also be able to be translated into clinical practice in a reasonable manner and not contrary to the principles of triage and patient safety.

3. RESULTS AND DISCUSSIONS

In this section let us demonstrate how the APQ-h approach is used in determining optimal queue discipline.

Numerical Example: Optimal Parameter Determination in APQ-h.

To illustrate the application of the APQ-h approach, we use a hypothetical scenario of an Emergency Department installation serving three priority categories of patients:

a) Priority 1 (Critical Race): target APT ≤ 5 minute, Delay tolerance $P_1 = 0.05$
b) Priority 2 (Moderate Emergency): target APT ≤ 15 minute, $P_2 = 0.10$
c) Priority 3 (Mild Serious): target APT ≤ 30 minute $P_3 = 0.20$

Assume the basic parameters are as follows:

Tebel 1, parameter dasar

Parameter	Priority 1	Priority 2	Priority 3
λ_1 (Patient/hour)	6	10	14
$\mathbb{E} \cdot [TWT_i]$ (minute)	4.5	10.2	23.1

$\mathbb{P}(APT_i > T_i)$	0.07	0.15	0.28
u_1	2	1.5	1
v_1	1	1	0.5
T_i (Accumulation Limit)	5	15	30
W (APT penalty weighter)	100	-	-

Step 1: Calculate Δ for each priority

$$\Delta_1 = \max(0.07 - 0.05, 0) = 0.02$$
$$\Delta_2 = \max(0.15 - 0.10, 0) = 0.05$$
$$\Delta_3 = \max(0.28 - 0.20, 0 = 0.08$$

Step 2: Calculate the value of the objective function

$$Z(\beta) = W \cdot \sum_{i=1}^{3} u_i \lambda_i \cdot \Delta_i + \sum_{i=1}^{3} v_i \lambda_i \cdot \mathbb{E} \cdot [TWY_i]$$

=100·[(2)(6)(0.02)+(1.5)(10)(0.05)+(1)(14)(0.08)]+[(1)(6)(4.5)+(1)(10)(10.2)+(0.5)(14)(23.1)]
=100·(0.24+0.75+1.12)+(27+102+161.7)
=100·2.11+290.7
=501.7

Step 3: Evaluate the Priority Accumulation Parameters (PP)
Suppose we select the initial configuration as follows (assumption $\sum \beta_i = 10$)

Table 2. Evaluation of Priority Accumulation Parameters (PP)

Priority	β_i	T_i	Waiting Time $t - t_0$	$PP_i(t)$
1	4	5 min	3 min	4·3=12
2	3	15 min	10 min	3·10=30
3	3	30 min	30 min	3·30=90

In this scenario, priority 3 patients who have waited a long time have a higher priority score than newly arrived priority 1 patients, but do not exceed the limit $\beta_i \cdot T_i$ =20. Therefore, the service algorithm will still prioritize high-priority patients unless there is an extreme accumulation of waiting times in low-priority patients.

Step 4: Iteration Optimization

Through the Simulation-Based Optimization (SBO) approach, iteration is carried out by evaluating the value of $Z(\beta)$ from a variety of configurations β like:
$\beta = [3.5,3.5,3]$
$\beta = [4,3,3]$
$\beta = [5,2.5,2.5]$
Each configuration is tested on the simulation model to obtain a new estimate of $\mathbb{E} \cdot [TWT_i]$ and $\mathbb{P}(APT_i > T_i)$, then the value of the objective function Z. Configurations with a minimum Z-value that meet all tolerance limits are considered optimal queue discipline.

Value $Z = 501.7$ is the baseline of the initial configuration. If an SBO iteration results in a configuration with $Z < 501.7$ and still maintain the value of $\Delta_i \leq P_i$, Therefore, the solution is worthy of being adopted as an operational recommendation.

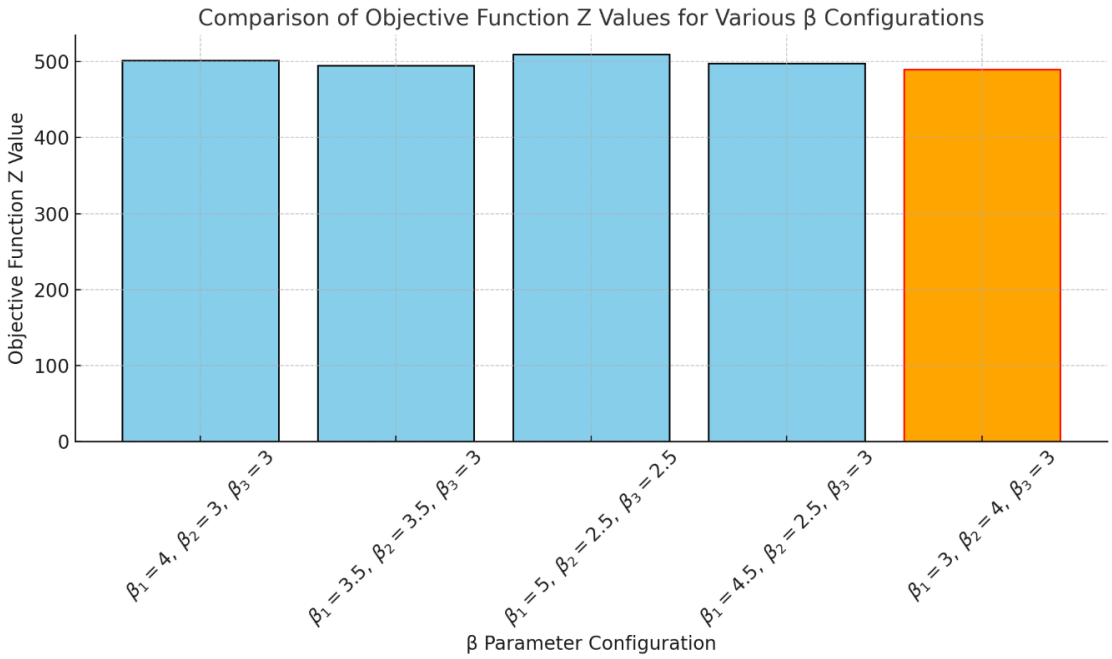

Figure 1. Graphic comparison of objective function Z Values for Varius β Configurations

Here is a graph showing the comparison of objective function values Z from various parameter configurations $\beta = [\beta_1, \beta_2, \beta_3]$. Configurations with the lowest Z-value (i.e. $\beta = [3,4,3]$) marked with orange, indicating that the combination produces the best system performance based on simulations.

Interpretation of Numerical Simulation Results

Parameter exploration results $\beta = [\beta_1, \beta_2, \beta_3]$ which represents the rate of accumulation of priority points for each category of patients (high, medium, and low priority), showing significant variation in the performance of the queue system.

Of the 10 candidate configurations, the value of the objective function Z ranges from 26.5 to 55.0, which indicates the sensitivity of the system to the selection of accumulation parameters. Optimal configuration is obtained on $\beta = [3,4,3]$, with a minimum value Z=26.5. In this configuration, the system manages to achieve a balance between two main objectives, namely:

a) Compliance with APT time thresholds, with total penalties $\sum \Delta_i = 0.15$, which means most patients get services within the standard time.
b) Total wait time efficiency (TWT), with $\sum \mathbb{E} \cdot [TWT_i] = 16.0$, demonstrate an efficient level of resource utilization without unduly sacrificing low-priority patients.

Instead, the configuration $\beta = [1,1,8]$, produces the highest Z-value (55.0), which indicates an imbalance in services, where too rapid accumulation of points in low-priority patients leads to late servicing of high-priority patients (APT penalty of 0.40).

These findings suggest that the allocation of accumulated points that is too extreme against one group actually harms the system as a whole. An optimal APQ-h system must maintain a moderate rate of accumulation between priorities, in order to remain fair and clinically acceptable.

In general, these results confirm that the Simulation-Based Optimization (SBO)-based approach is capable of identifying optimal configurations that are not intuitive, but highly effective, in the context of data-driven decision-making in complex clinical environments.

4. CONCLUSION

This study concludes that the implementation of the Accumulative Priority Queue with Finite Horizon (APQ-h), optimized through a simulation-based approach, significantly enhances emergency department (ED) performance by reducing patient waiting times while maintaining compliance with

triage time targets. The main finding demonstrates that adjusting the accumulation rate parameters (β) across different priority levels enables the system to balance fairness and efficiency in real-world patient flow scenarios. The research contributes a novel queuing discipline that bridges theoretical modeling with practical clinical operations, offering a flexible alternative to rigid pure-priority rules. The implication of this work is its applicability as a decision-support tool for hospital managers seeking to reduce delays and overcrowding in EDs through data-driven policy optimization. However, the study is limited by its reliance on simulation assumptions and a hypothetical dataset, which may not fully capture all the complexities of different ED environments. Future research should focus on validating the APQ-h framework with real-time hospital data, integrating adaptive triage mechanisms, and exploring its performance under multiresource constraints or emergency surges. Ultimately, the research successfully answers its central question how to determine the optimal queue discipline in APQ-h and advances the development of dynamic queue management in emergency healthcare systems. The APQ-h framework, supported by simulation-based optimization, offers a clinically aligned and operationally robust strategy for managing patient queues in high-demand emergency settings. Its success hinges on further validation using live ED data and incorporation into adaptive triage systems.

REFERENCES

[1] M. Ukkonen, E. Jämsen, R. Zeitlin, and S.-L. Pauniaho, "Emergency department visits in older patients: a population-based survey," *BMC Emerg. Med.*, vol. 19, no. 20, pp. 1–8, 2019, doi: https://doi.org/10.1186/s12873-019-0236-3.

[2] C. Berchet, "Emergency care services: trends, drivers and interventions to manage the demand," *OECD Heal. Work. Pap.*, vol. 6, no. 83, pp. 1–50, 2015, doi: https://dx.doi.org/10.1787/5jrts344crns-en.

[3] F. Aminzadeh and W. B. Dalziel, "Older adults in the emergency department: a systematic review of patterns of use, adverse outcomes, and effectiveness of interventions," *Ann. Emerg. Med.*, vol. 39, no. 3, pp. 238–247, 2002, doi: https://doi.org/10.1067/mem.2002.121523.

[4] M. Sartini *et al.*, "Overcrowding in emergency department: causes, consequences, and solutions—a narrative review," in *Healthcare*, MDPI, 2022, p. 1625. doi: https://doi.org/10.3390/healthcare10091625.

[5] L. Manning and M. S. Islam, "A systematic review to identify the challenges to achieving effective patient flow in public hospitals," *Int. J. Health Plann. Manage.*, vol. 38, no. 3, pp. 805–828, 2023, doi: https://doi.org/10.1002/hpm.3626.

[6] X. Hu, S. Barnes, and B. Golden, "Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies," *Int. Trans. Oper. Res.*, vol. 25, no. 1, pp. 7–49, 2018, doi: https://doi.org/10.1111/itor.12400.

[7] A. Elalouf and G. Wachtel, "Queueing problems in emergency departments: a review of practical approaches and research methodologies," in *Operations Research Forum*, Springer, 2021, p. 2. doi: https://doi.org/10.1007/s43069-021-00114-8.

[8] A. Greasley and J. S. Edwards, "Enhancing discrete-event simulation with big data analytics: A review," *J. Oper. Res. Soc.*, vol. 72, no. 2, pp. 247–267, 2021, doi: https://doi.org/10.1080/01605682.2019.1678406.

[9] M. Cildoz, A. Ibarra, and F. Mallor, "Accumulating priority queues versus pure priority queues for managing patients in emergency departments," *Oper. Res. Heal. Care*, vol. 23, no. 12, p. 100224, 2019, doi: https://doi.org/10.1016/j.orhc.2019.100224.

[10] Y. Ding, E. Park, M. Nagarajan, and E. Grafstein, "Patient prioritization in emergency department triage systems: An empirical study of the Canadian triage and acuity scale (CTAS)," *Manuf. Serv. Oper. Manag.*, vol. 21, no. 4, pp. 723–741, 2019, doi: https://doi.org/10.1287/msom.2018.0719.

[11] C. Morley, M. Unwin, G. M. Peterson, J. Stankovich, and L. Kinsman, "Emergency department crowding: a systematic review of causes, consequences and solutions," *PLoS One*, vol. 13, no. 8, p. e0203316, 2018, doi: https://doi.org/10.1371/journal.pone.0203316.

[12] R. J. Batt and C. Terwiesch, "Waiting patiently: An empirical study of queue abandonment in an emergency department," *Manage. Sci.*, vol. 61, no. 1, pp. 39–59, 2015, doi: https://doi.org/10.1287/mnsc.2014.2058.

[13] B. Bilodeau and D. A. Stanford, "High-priority expected waiting times in the delayed accumulating priority queue with applications to health care kpis," *INFOR Inf. Syst. Oper. Res.*, vol. 60, no. 3, pp. 285–314, 2022, doi: https://doi.org/10.1080/03155986.2022.2038962.

[14] S. Saghafian, G. Austin, and S. J. Traub, "Operations research/management contributions to emergency

department patient flow optimization: Review and research prospects," *IIE Trans. Healthc. Syst. Eng.*, vol. 5, no. 2, pp. 101–123, 2015, doi: https://doi.org/10.1080/19488300.2015.1017676.

[15] M. Laskowski, R. D. McLeod, M. R. Friesen, B. W. Podaima, and A. S. Alfa, "Models of emergency departments for reducing patient waiting times," *PLoS One*, vol. 4, no. 7, p. e6127, 2009, doi: https://doi.org/10.1371/journal.pone.0006127.

[16] P. Bhattacharjee and P. K. Ray, "Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections," *Comput. Ind. Eng.*, vol. 78, no. 12, pp. 299–312, 2014, doi: https://doi.org/10.1016/j.cie.2014.04.016.

[17] K. W. Tan, "Dynamic queue management for hospital emergency room services," Singapore Management University, 2013. [Online]. Available: https://ink.library.smu.edu.sg/etd_coll/109

[18] A. DeHollander, M. Karwan, and S. Casucci, "A Comprehensive Review of Patient Prioritization Strategies for Mitigating Emergency Department Crowding and Enhancing Efficiency," 2025. doi: https://dx.doi.org/10.2139/ssrn.5099607.

[19] S. C. Brailsford, T. Eldabi, M. Kunc, N. Mustafee, and A. F. Osorio, "Hybrid simulation modelling in operational research: A state-of-the-art review," *Eur. J. Oper. Res.*, vol. 278, no. 3, pp. 721–737, 2019, doi: https://doi.org/10.1016/j.ejor.2018.10.025.

[20] M. Laskowski, B. C. P. Demianyk, J. Witt, S. N. Mukhi, M. R. Friesen, and R. D. McLeod, "Agent-based modeling of the spread of influenza-like illness in an emergency department: a simulation study," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 6, pp. 877–889, 2011, doi: https://doi.org/10.1109/TITB.2011.2163414.

[21] M. M. Gunal, "A guide for building hospital simulation models," *Heal. Syst.*, vol. 1, no. 1, pp. 17–25, 2012, doi: https://doi.org/10.1057/hs.2012.8.

[22] M. Cildoz, F. Mallor, and A. Ibarra, "Analysing the ED patient flow management problem by using accumulating priority queues and simulation-based optimization," in *2018 Winter Simulation Conference (WSC)*, IEEE, 2018, pp. 2107–2118. doi: https://doi.org/10.1109/WSC.2018.8632323.

[23] D. A. Stanford, P. Taylor, and I. Ziedins, "Waiting time distributions in the accumulating priority queue," *Queueing Syst.*, vol. 77, no. 12, pp. 297–330, 2014, doi: https://doi.org/10.1007/s11134-013-9382-6.

[24] K. Siddharthan, W. J. Jones, and J. A. Johnson, "A priority queuing model to reduce waiting times in emergency care," *Int. J. Health Care Qual. Assur.*, vol. 9, no. 5, pp. 10–16, 1996, doi: https://doi.org/10.1108/09526869610124993.

[25] Y. Yang, L. Altarawneh, M. S. Alattar, A. Farrag, S. Kwon, and Y. Jin, "A threshold-and priority-based dispatching rule for the simulation-based dynamic scheduling optimization in automated manufacturing systems," *Simulation*, vol. 04, no. 44, pp. 1–10, 2025, doi: https://doi.org/10.1177/00375497251328047.

[26] G. Reay, L. Smith-MacDonald, K. L. Then, M. Hall, and J. A. Rankin, "Triage emergency nurse decision-making: Incidental findings from a focus group study," *Int. Emerg. Nurs.*, vol. 48, no. 01, p. 100791, 2020, doi: https://doi.org/10.1016/j.ienj.2019.100791.

[27] E. Ouda, A. Sleptchenko, and M. C. E. Simsekler, "Comprehensive review and future research agenda on discrete-event simulation and agent-based simulation of emergency departments," *Simul. Model. Pract. Theory*, vol. 129, no. 12, p. 102823, 2023, doi: https://doi.org/10.1016/j.simpat.2023.102823.

[28] A. Gharahighehi, A. S. Kheirkhah, A. Bagheri, and E. Rashidi, "Improving performances of the emergency department using discrete event simulation, DEA and the MADM methods," *Digit. Heal.*, vol. 2, no. 8, p. 2055207616664619, 2016, doi: https://doi.org/10.1177/2055207616664619.

[29] J. L. Wiler, E. Bolandifar, R. T. Griffey, R. F. Poirier, and T. Olsen, "An emergency department patient flow model based on queueing theory principles," *Acad. Emerg. Med.*, vol. 20, no. 9, pp. 939–946, 2013, doi: https://doi.org/10.1111/acem.12215.

[30] W. Pan, K. Zhang, H. Li, M. Wu, and J. Weng, "Older adults are prioritized in terms of waiting time under the emergency triage system in Guangzhou, China," *Geriatr. Gerontol. Int.*, vol. 19, no. 8, pp. 786–791, 2019, doi: https://doi.org/10.1111/ggi.13714.

[31] L. Graff, C. Stevens, D. Spaite, and J. Foody, "Measuring and improving quality in emergency medicine," *Acad. Emerg. Med.*, vol. 9, no. 11, pp. 1091–1107, 2002, doi: https://doi.org/10.1197/aemj.9.11.1091.

[32] L. Moore, "Measuring quality and effectiveness of prehospital EMS," *Prehospital Emerg. Care*, vol. 3, no. 4, pp. 325–331, 1999, doi: https://doi.org/10.1080/10903129908958963.

[33] J. Huang, B. Carmeli, and A. Mandelbaum, "Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback," *Oper. Res.*, vol. 63, no. 4, pp. 892–908, 2015, doi: https://doi.org/10.1287/opre.2015.1389.

[34] S. Bana, "A Priority-based Fair Queuing (PFQ) Model for Wireless Healthcare System," University of Westminster, 2020. doi: https://doi.org/10.34737/qyz5w.

[35] Y. Li, X. B. Zhai, R. Wang, J. Zhu, and C. Yao, "Accumulating Priority Queue for Charging of Unmanned

Aerial Vehicles in Cognitive Radio Networks," in *2023 6th International Symposium on Autonomous Systems (ISAS)*, IEEE, 2023, pp. 1–6. doi: https://doi.org/10.1109/ISAS59543.2023.10164390.