ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202206.00069

Analysis of Affective and Rational Reneging Rates in Dynamic Overflow Queuing Systems for Outpatient Services (Postprint)

Authors: Li Junxiang, Lu Yiling, Longbo Pan

Date: 2022-06-06T00:00:00+00:00

Abstract

In real-world healthcare delivery, the traditional dual-queue independent queuing system struggles to efficiently utilize hospital resources amid the prevailing shortage of physician capacity. To address this challenge, we propose establishing system dynamics to rationally allocate patient volumes across outpatient departments without modifying existing hospital resources. Building upon the conventional dual-queue independent queuing framework, we develop a dynamic overflow outpatient queuing birth-death model that incorporates patient departure probabilities driven by both emotional and rational factors, thereby enabling more accurate analysis of the system' s true state. Leveraging ProModel —a flexible and reliable discrete-event simulation software—we conduct visualized model studies and comparative analyses against traditional outpatient queuing models. The simulation results indicate that the novel queuing model yields superior performance across key service metrics, including average queue length and patient waiting times.

Full Text

Preamble

Analysis of Exit Rate with Perceptualual and Rational Factors in Outpatient Dynamic Overflow Queuing System

Li Junxiang†, Lu Yiling, Pan Longbo
(Business School, University of Shanghai for Science & Technology, Shanghai 200093, China)

Abstract: In practical medical service applications, the traditional double-queue independent queuing system struggles to efficiently utilize hospital resources under conditions where doctor capacity is in short supply. To address

this challenge without altering existing hospital resources, this study establishes a dynamic overflow outpatient queuing birth-death model based on the traditional double-queue independent system. The model incorporates patient exit probabilities under both perceptual and rational factors to enable more accurate analysis of the system' s real state. Using ProModel, a flexible and reliable discrete-event simulation software, the model is visualized and compared with traditional outpatient queuing models. Simulation results demonstrate that the new queuing model achieves superior performance across service metrics including average queue length and patient waiting time.

Keywords: queuing system; dynamic overflow; exit rate; perceptual; rational

0 Introduction

In recent years, with the continuous improvement of medical service levels and expanding public demand, China' s healthcare sector has developed rapidly. Reports from the National Health Commission have emphasized that a nation' s medical quality and safety directly impact population health [1]. To build a high-quality, efficient healthcare service system that better satisfies growing health needs, countries have significantly improved medical technology capabilities and quality levels while steadily increasing healthcare investments. However, despite substantial increases in healthcare funding and certain achievements in medical system reform, most hospitals have not achieved proportional improvements in operational efficiency. They continue to face critical issues including high outpatient volumes, crowded patient queues, and excessively long waiting times. Data indicates that effective consultation time accounts for only 10%-15% of total outpatient time, while non-treatment time reaches approximately 85% [2]. This demonstrates that ensuring medical quality and patient safety requires not only financial investment but also investigation into how to effectively improve the efficiency of various operational segments within healthcare service systems.

Numerous scholars have conducted research on improving hospital operational efficiency. Joseph et al. [3] employed retrospective cohort studies using generalized estimating equations to predict hourly patient volumes per physician and quantified relationships between physician workload, patient safety, department location, and service quality. Andersen et al. [4] modeled patient flow as transient queuing networks, using integer programming recursion to evaluate and allocate patients until all queues met waiting time targets. Zhang et al. [5] proposed a novel priority-based patient queuing model to optimize emergency service management, considering key service factors and exploring emergency service optimization through relevant queuing metrics compared to classical queuing rules. Barros et al. [6] analyzed prediction errors between demand and available medical resources, using simulation models to evaluate performance differences across various facility and resource configurations. Wu et al. [7] applied multi-stage tandem queuing models with bed capacity and

service interactions to model healthcare systems. Li Lingyang et al. [8] studied hospital window optimization using M/M/c/∞/∞/FCFS queuing models based on research into stochastic dynamic characteristics of hospital charging processes. Jing Tong et al. [9] examined retrial queuing systems with working breakdowns and customer balking, providing risk prediction and decision evaluation for realistic hospital operations. Yang Ying et al. [10] established a value model for patient decisions on selecting specialists, considering factors such as specialist service quality, patient queuing costs, and anxiety levels. Using this model, they derived performance metrics including maximum acceptable waiting times and queue lengths, proposing detailed optimization recommendations. He et al. [11] developed a dental service process optimization model for a tertiary hospital in Changzhou using discrete-event simulation, collecting data through field research and setting model parameters based on existing service models. They recommended adding pre-examination to triage processes to improve service quality and resource utilization while reducing patient waiting time under limited resources. Kang et al. [12] utilized Korea' s National Emergency Medical Information System and artificial intelligence techniques including feedforward neural networks and regularization to analyze data from multiple emergency patients meeting standards, establishing an emergency severity triage algorithm that outperformed commonly used clinical triage algorithms.

In summary, existing literature primarily focuses on independent queue static optimization across dimensions including medical staff allocation, patient decision-making, and key service metrics, employing queuing theory and algorithmic simulation for effective solutions. However, in practical scenarios, hospital triage allocation and dynamic queue scheduling optimization can substantially improve service operational efficiency. Current research in this area remains limited, with numerous studies [11,12] remaining at the theoretical discussion level.

To address these research gaps, this paper proposes a novel triage model that incorporates patient psychological and rational factors, examining queuing system states from perspectives of physician service efficiency and patient overflow operations while conducting quantitative research on various system metrics. First, based on the traditional double-queue independent queuing system, system dynamics are added: when one queue reaches capacity while another has waiting space, nurses can perform secondary triage to overflow patients from the saturated queue to the alternative queue based on patient willingness. Second, the probability of patients exiting the system due to psychological and rational factors is analyzed, comprehensively considering multiple constraints to maintain dynamic balance between reducing patient waiting time and not increasing exit rates. Finally, the model is compared with traditional outpatient queuing models without changing existing hospital resources, achieving research objectives of reducing average queue length and patient waiting time. Due to strong model variability, traditional algorithms or equilibrium state equations yield static average values [13], whereas this study' s queuing system represents a continuously changing dynamic process that cannot be reflected through fixed

values. Therefore, ProModel simulation is employed for system modeling, generating variation graphs of relevant performance metrics during operation. By observing peaks, valleys, and trend changes, the service process can be more effectively and intuitively demonstrated.

---

1 Model Description

This section describes the outpatient overflow queuing system considering two-way patient overflow and establishes relevant parameters. Patients arrive at the hospital according to a Poisson process with parameter λ_i ($i = 1, 2$ representing general and specialist patients respectively). The system has c physician service seats, with c_i representing the number of physician seats in queue i. When patients arrive and find all physician seats busy, they must join the queue for service.

If waiting queues become excessively long or physician service efficiency is low, patients may become emotionally agitated and even choose to leave the queuing system. When no queue waiting is required, the probability of patients entering the system is 1. When queue waiting numbers exceed 0, not all patients will join the waiting queue. Patient exit factors arise from two situations: first, psychological utility—patients decide whether to exit based on the difference between service utility received and costs incurred. If they choose to exit, some patient loss occurs at the initial system entry stage, with the assumed queue entry rate parameter being λ_i'; second, rational utility—during queuing, patients decide whether to exit based on queue length and individual sensitivity parameters.

When queue i reaches its saturation threshold upper limit $L_i^{\max}$ and the other queue has not reached its saturation threshold lower limit $L_i^{\min}$, the system intervenes and determines whether overflow triage conditions are met. If conditions are satisfied, patients overflow to the other outpatient system at rate λ_i'. Otherwise, patients continue waiting in the current queue. To ensure hospital operational benefits and meet patient consultation needs while guaranteeing the overflow system does not lose more patients, the system requires that the number of exiting patients from each queue not exceed k times the overflow population.

When physician seats operate normally, the service rate for queue i follows an exponential distribution with parameter μ_i. During patient treatment, physicians adjust their average service intensity based on working hours. When the total patient threshold in the system at time t is n_i' (for calculation convenience, the total patient threshold at this time is denoted as n_i', representing the total patient threshold for queue i at this moment), the physician service rate is $\mu_{i,1}$ (representing the average number of patients treated per unit time), where $\rho_{i,1}$ is the treatment intensity under fast treatment rates. Conversely, when the number of patients in queue i is less than n_i', service rates decrease due

to fatigue and are set as $\mu_{i,2}$, where $\rho_{i,2}$ is the treatment intensity under slow treatment rates, with $\mu_{i,1} > \mu_{i,2}$.

Patient arrival intervals, physician seat service rates, and patient impatience intensity are mutually independent. The system service discipline follows the First-In-First-Out (FIFO) principle.

Considering these factors, the routing rules for the secondary triage consultation queuing model are established as shown in [Figure 1: see original paper]. Patients register through multiple channels (appointment or on-site), with registration information guiding them to the nurse station in the corresponding clinic area for consultation confirmation. The triage waiting unit in the clinic area enables multi-functional business management and operations. Patients are divided into general and specialist outpatient categories based on registration requirements. First, psychological utility is considered: the system determines whether patient service utility exceeds the sum of service fees and corresponding waiting costs. If satisfied, patient psychological utility is positive and they can enter the appropriate queue; otherwise, they exit the system. Second, overflow conditions are assessed: when a queue reaches saturation, the system intervenes and determines whether overflow triage conditions are met. If satisfied, patients can overflow to the other outpatient system; otherwise, they continue waiting in the current queue. Finally, patients decide whether to exit based on rational utility: after entering the queue, patients determine whether to exit due to excessive waiting time caused by decreased physician efficiency or overflow operations; otherwise, they continue waiting for service. While patients wait for service using FIFO rules, this routing reduces crowding in waiting areas and improves utilization of outpatient physician seats.

2.1 Objective Function

This study examines an M/M/c/N consultation queuing model with limited service capacity and variable service rates, analyzing state transitions within the system. This model represents a capacity-limited multi-server system where the queuing system state can be represented by elements from the following set [14]:

$$S(t) = \{(n_1(t), n_2(t)) \mid 0 \leq n_1(t) < N_1, 0 \leq n_2(t) \leq N_2\}$$

where $n_i(t)$ represents the number of patients in queue i at time t, and T is the total working time of outpatient physicians. N_i is the maximum capacity of queue i in the outpatient system.

Research results from relevant literature [15] indicate that physician service rates are influenced by factors including working hours, number of patients treated, and patient condition complexity. Assuming service rates vary with the number

of patients in the system, the physician service rate for queue i is given by equation (2):

$$\mu_i(n_i(t)) = \begin{cases} \mu_i & 0 < n_i(t) < c_i \\ \mu_i & c_i \le n_i(t) < c_i + n_i' \\ \mu_{i,2} & n_i(t) \ge c_i + n_i' \end{cases}$$

A birth-death process is employed to model this scenario, where finding steady-state conditions represents an important problem. This section presents theorems and proofs for steady-state conditions of the birth-death process. Let $P_i(n,t)$ denote the probability that n patients have arrived in queue i by time t, with $P_i(0)$ representing the initial state probability of no patients in queue i.

Theorem: If $\rho_{i,1} = \frac{\lambda_i}{\mu_{i,1}} < 1$ and $\rho_{i,2} = \frac{\lambda_i}{\mu_{i,2}} < 1$ (i.e., the system is stable), then a steady-state distribution exists for the queuing system with variable service rates:

$$P_i(n) = \begin{cases} \frac{\rho_i^n}{n!} P_0 & 0 \le n < c_i \\ \frac{\rho_i^{c_i} \rho_{i,1}^{n-c_i}}{c_i!} P_0 & c_i \le n < c_i + n_i' \\ \frac{\rho_i^{c_i} \rho_{i,1}^{n_i'} \rho_{i,2}^{n-c_i-n_i'}}{c_i!} P_0 & n \ge c_i + n_i' \end{cases}$$

Proof: From the Kolmogorov algebraic equations for birth-death process equilibrium (where input rate equals output rate), we derive:

For $n = 0$: $P_{i,0} = \frac{\lambda_i}{\mu_i} P_{i,0}$

For $1 \le n < c_i$: $P_{i,n} = \frac{\lambda_i}{n\mu_i} P_{i,n-1} = \frac{\rho_i^n}{n!} P_{i,0}$

For $n = c_i - 1$: $c_i \mu_i P_{i,c_i} = \lambda_i P_{i,c_i-1}$

For $c_i \le n < c_i + n_i'$: $P_{i,n} = \frac{\lambda_i}{c_i \mu_{i,1}} P_{i,n-1} = \frac{\rho_i^{c_i} \rho_{i,1}^{n-c_i}}{c_i!} P_{i,0}$

For $n = c_i + n_i'$: $c_i \mu_{i,2} P_{i,c_i+n_i'} = \lambda_i P_{i,c_i+n_i'-1}$

For $n \ge c_i + n_i'$: $P_{i,n} = \frac{\lambda_i}{c_i \mu_{i,2}} P_{i,n-1} = \frac{\rho_i^{c_i} \rho_{i,1}^{n_i'} \rho_{i,2}^{n-c_i-n_i'}}{c_i!} P_{i,0}$

After obtaining the steady-state probabilities $P_i(n,t)$, the system' s main performance indicators require calculation as follows:

In an M/M/c/N system, waiting patients indicate that patient numbers exceed physician seats, meaning c_i patients are receiving service in steady state. According to the mean definition [16], the average queue length $L_i(t)$ for queue i at time t is:

$$L_i(t) = \sum_{n=c_i+1}^{N_i} (n - c_i) P_i(n, t)$$

By Little' s Law [17], the average waiting time $W_i(t)$ for queue i at time t is the ratio of waiting patients' average queue length to patient arrival rate. Let $W_i'(t)$ represent the average waiting time in the traditional queuing model. The total patient waiting time in the traditional double-queue independent system is:

$$\Gamma_1(t) = \sum_{i=1}^{2} \sum_{t=1}^{T} L_i(t) W_i'(t)$$

Let $W_i(t)$ represent the average waiting time in the new overflow queuing model. The total patient waiting time in this system is:

$$\Gamma_2(t) = \sum_{i=1}^{2} \sum_{t=1}^{T} L_i(t) W_i(t)$$

2.2 Constraints

1) Psychological Factor Exit Constraint

Utility typically measures a service' s ability to satisfy patient needs and represents an important evaluation metric for patient decision-making on whether to accept service. Based on utility theory [18], patient psychological utility is defined as the psychological satisfaction degree obtained from physician services. Assume patients obtain treatment effectiveness R_i from physician services, pay service fees C_i, and incur corresponding unit time waiting costs K_i.

The psychological utility $U_i(t)$ for general and specialist patients is:

$$U_1(t) = [R_1 - C_1 - K_{1E}[W_1(t)] - E[W_2(t)]]\sigma_1$$
$$U_2(t) = [R_2 - C_2 - K_{2E}[W_2(t)] - E[W_1(t)]]\sigma_2$$

where σ_i represents patient preference values, positively correlated with expected service utility at time t. Since specialist outpatient service fees exceed general outpatient fees ($C_2 > C_1$), patient service utility shows strong correlation. Patient decisions to enter the service system depend on psychological utility, exiting when $U_i(t) \leq 0$. Therefore, the unit waiting time cost for exiting patients should satisfy:

$$K_i \geq \frac{R_i - C_i}{E[W_i(t)]}$$

Assume the output rates of general and specialist patients exiting due to psychological utility at time t are $P_1(t)$ and $P_2(t)$ respectively. From the above analysis, the patient arrival rate entering queues at time t is $\lambda_i' = \lambda_i(1 - P_i(t))$.

2) Patient Overflow Operation Queue Constraints

Determining whether patients choose overflow operations requires satisfying two conditions:

a) **Queue Constraints**: When the average queue length of waiting patients in queue i reaches threshold upper limit $L_i^{\max}$ and the other queue' s average queue length remains within threshold lower limit $L_i^{\min}$, the system can provide overflow options. Queue constraints are:

$$\gamma_{i,1}(t) = \begin{cases} 1 & \text{if } L_1(t) \geq L_1^{\max} \text{ and } L_2(t) \leq L_2^{\min} \\ 0 & \text{otherwise} \end{cases}$$

$$\gamma_{i,2}(t) = \begin{cases} 1 & \text{if } L_2(t) \geq L_2^{\max} \text{ and } L_1(t) \leq L_1^{\min} \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma_{i,j}(t)$ are 0-1 variables. When $\gamma_{1,1}(t) = 1$ and $\gamma_{1,2}(t) = 0$, this represents general patient overflow queue constraints. When $\gamma_{2,2}(t) = 1$ and $\gamma_{2,1}(t) = 0$, this represents specialist patient overflow queue constraints.

b) **Patient Willingness Decision Constraint**: Based on Markowitz' s mean-variance theory [19], patient decision constraints for queue transfer are established using the trade-off relationship between average waiting time $W_i(t)$ and variance as important decision indicators. When patients objectively weigh options, asymmetric preferences exist between mean and variance. Research shows [20] that patients prefer mean reduction over standard deviation reduction. Assume ν $(0 < \nu < 1)$ is the patient' s asymmetric preference coefficient, and $\pi_{i,t}$ represents general patient willingness to overflow at time t. When both the mean and standard deviation of general patient waiting time exceed those of the specialist queue, or when mean reduction exceeds ν times the standard deviation increase, $\pi_{i,t} = 1$; otherwise $\pi_{i,t} = 0$.

The overflow system' s average queue length is $L_i'(t) = L_i(t) + \Delta L_i(t)$, where $\Delta L_i(t)$ represents the overflow patient count.

3) Rational Factor Exit Constraint

While implementing overflow operations, patient impatience factors are considered. Excessively long waiting times may cause 烦躁情绪 (agitation) and lead patients to leave the system. Assume the intensity of impatient patients leaving the system at time t is linearly related to the average queue length $L_i(t)$ in steady state and the overflow system' s average queue length $L_i'(t)$. The patient exit probability is $\Delta\varepsilon_i(t) = \frac{L_i'(t) - L_i(t)}{L_i(t)}$. Patient choices are independent, with

individual tolerance levels varying. Assume patient sensitivity parameter ε is uniformly distributed on $[\underline{\varepsilon}, \overline{\varepsilon}]$, where $\underline{\varepsilon}$ and $\overline{\varepsilon}$ represent lower and upper bounds of patient sensitivity parameters. A smaller ε indicates higher sensitivity and greater impatience.

The patient exit rate is a linear function of average queue length, with balking probability [21]:

$$\Delta\varepsilon_i(t) = \frac{L_i'(t)}{L_i(t)} - 1$$

The total number of general and specialist patients exiting the system at time t is $l_i(t) = \Delta\varepsilon_i(t)L_i'(t)$. To balance physician seat utilization without causing massive patient loss, the number of patients leaving queues at time t must not exceed k times the overflow patient count:

$$l_i(t) \leq k\Delta L_i(t)$$

This study aims to maximize the difference in total waiting time for all patients before and after overflow operations, indicating the overflow queuing system can significantly reduce patient waiting time. Integrating all constraints, the optimization model is:

$$\max \Delta\Gamma = \Gamma_1(t) - \Gamma_2(t)$$

Subject to:

$$\begin{cases} \lambda_i' = \lambda_i(1 - P_i(t)) \\ \gamma_{i,j}(t) \in \{0,1\}, \gamma_{i,1}(t) \neq \gamma_{i,2}(t) \\ L_i'(t) = \sum_{t=1}^{T} \pi_{i,t} \\ \Delta\varepsilon_i(t) = \frac{L_i'(t)}{L_i(t)} - 1 \\ l_i(t) = \Delta\varepsilon_i(t)L_i'(t) \leq k\Delta L_i(t) \end{cases}$$

3.1 Simulation Module Analysis

ProModel is a discrete-event system simulation software with time resolution ranging from 0.01 hours to 0.0001 seconds, enabling high simulation precision [22]. It is primarily applied to production and service system simulation. This study uses ProModel for simulation modeling to reduce patient waiting time and optimize outpatient service metrics.

Based on realistic scenarios, experimental and control groups are established for comparative analysis. The experimental group incorporates the overflow queuing model with its three constraints (as shown in [Figure 2: see original

paper]), with other variables consistent with the control group to investigate whether the new model impacts service metrics.

ProModel modeling elements include system object elements and system operations, involving over a dozen module types. Basic element parameter settings are shown in Tables 1-3.

1) Location/Entities

Location represents fixed places or areas where patients wait in the hospital, including corresponding capacity and unit numbers, with time series set to collect basic data and track location time series values. Entities are the objects served in the simulation, referring to patients in the hospital (using both "patients" and "clients" to distinguish entities at different locations).

2) Arrival

Patient arrivals have capacity limits and follow specific probability distributions. To enhance practical value, field research data from a hospital' s surgical department was processed as simulation parameters. Based on one working day' s data: general outpatient hours are 8:00-16:15; specialist outpatient hours are 8:00-12:00 and 13:00-16:00. Using SPSS for one-sample Kolmogorov-Smirnov tests on arrival data and service times for both general and specialist patients, results show asymptotic significance (two-tailed) > 0.05, indicating Poisson distribution for arrival rates and exponential distribution for service times. Both patient arrival rates λ_i and queue entry rates λ_i' follow Poisson distributions. The Arrivals interface settings are shown in , where the system prohibits entry when patient numbers exceed capacity limits.

3) Cost

The cost module constrains psychological utility exits to determine whether patients exit directly due to negative psychological utility. Based on literature [23] examining supply-demand interests, utility parameters are set: general and specialist patient service fees are 15 and 40 yuan respectively, with waiting costs of 0.5 and 0.65, as shown in .

4) Process

a) The module first reads physician work schedules from the Arrival module to determine each queue's maximum capacity and whether patient entities can enter respective systems.

b) General and specialist patient systems operate in parallel, with similar patient entities stored in the same queue, following FIFO rules while waiting for physician seats. In Process Order 1, entities receive system status information (psychological utility constraints) to determine whether to exit, with entities exiting at rate P(0.03). After decisions, remaining entities are tagged to 统计实际进入系统的泊松流 (stat the actual Poisson flow entering the system).

c) When tagged entities enter queue waiting areas, they trigger patient count statistics. Order 2 checks whether the queue exceeds storage threshold $L_i^{\max}$ while the other queue remains within storage threshold $L_i^{\min}$. Among entities exceeding the threshold, transfer willingness is assessed. Eligible entities transfer to the other queue's temporary area at frequency P(0.232) to continue waiting; otherwise, they proceed directly to Order 3.

d) Order 3 entities directly enter the queue' s waiting temporary area. Entities in each queue' s temporary area wait for physician seats. When average queue length and overflow system average queue length reach thresholds, entity waiting times are evaluated against upper limits to determine whether entities exit or continue waiting. Specific routing rules are shown in , with general patient processes listed as examples and specialist patients following identical logic.

3.2 Simulation Results Analysis

The experimental and simulation environment uses Windows 10 with LINGO 18.0 and ProModel version 7.0. To validate simulation effectiveness, LINGO solver is used for comparison with traditional queuing models. Since LINGO is suitable for linear/nonlinear optimization but cannot effectively quantify overflow model constraints, three different patient arrival rates are set (0.8×, 1×, and 1.2×) for horizontal comparison of average waiting time changes across time periods (only 120, 240, and 360-minute results are listed). Results in show average difference between theoretical and simulation values is 9.6444E-02, with average variance of 2.386E-03, indicating relatively accurate simulation data.

demonstrates that as arrival rates increase positively, average waiting time extends accordingly without abnormal test results, with trends conforming to reality. Horizontal comparison between traditional and optimized models shows optimized average waiting time exhibits clear advantages at the same time points under consistent arrival rates, indicating positive impacts from the optimized routing and constraints.

Note: In , the first row shows $\lambda_1 = 12.976$, $\lambda_2 = 8.249$, units in patients/hour.

Under controlled variables, traditional hospital queuing models and two-way overflow models considering psychological/rational exit factors were simulated for 480 minutes across 20 runs at 90% confidence level to ensure balanced stable states.

[Figure 3: see original paper] compares patient output numbers before and after optimization. Two conclusions emerge: First, the total output of general and specialist queues exceeds that of the traditional model, indicating significant improvement in patients served per unit time. Second, in traditional models, specialist queue output far exceeds general queue output, potentially causing

sharp increases in specialist queue blocking rates or general physician seat idleness. The optimized model achieves relatively balanced output between queues with gradually converging curves, effectively avoiding these issues.

[Figure 4: see original paper] compares patient exit numbers before and after optimization. Both traditional specialist and general queue exit numbers exceed optimized exit numbers (including total exits from psychological and rational factors). Two reasons explain this: First, the optimized model improves physician seat utilization, reducing patient waiting time at different stages and thus decreasing exits due to impatience or excessive waiting. Second, the optimized model provides multiple options and service channels rather than forcing patients to choose between exiting or waiting for the next time slot, offering more flexible and humanized solutions that reduce overall exit numbers.

[Figure 5: see original paper] shows overflow patient numbers in general and specialist queues within the new two-way overflow model. Results indicate that general-to-specialist overflow exceeds specialist-to-general overflow in real-time dynamics, suggesting most specialist registrations stem from medical necessity with patients believing specialist diagnosis is more accurate. Patients overflowing from specialist to general queues may have chosen specialists based on psychological needs rather than medical requirements. Conversely, when general queue patients discover longer waiting times compared to specialist queues, they are willing to pay higher fees to overflow to specialist queues for time savings.

[Figure 6: see original paper] compares average queue lengths before and after optimization. In traditional models, general and specialist queue lengths stabilize at approximately 30-40 and 10-20 patients respectively. After adding two-way overflow channels, both queues show significant length reductions. General queue length is controlled within 0-5 patients, while specialist queue length stabilizes at 0-6 patients, with occasional increases due to general-to-specialist overflow that remain within reasonable fluctuation ranges.

[Figure 7: see original paper] compares average waiting times. In traditional models, both general and specialist queue waiting times exhibit irregular violent fluctuations within 20-40 minutes, indicating poor system stability that may increase patient impatience and trigger conflicts. The optimized model controls waiting times within 10 minutes for all queues, showing significant reductions compared to traditional models with relatively stable curve fluctuations, demonstrating better system stability.

---

4 Conclusion

This paper explores a two-way overflow queuing model considering physician fatigue-induced service rate changes. First, patient consultation queuing problems and routing rules are described and analyzed. Second, queuing theory and quasi-birth-death processes quantify various evaluation indicators, with

constraints established for system capacity, patient decisions, and exit rates. Finally, ProModel simulation compares models and LINGO validates simulation effectiveness. Under controlled variables, the new model not only reduces patient waiting time throughout the consultation process but also reasonably balances patient-physician ratios and service intensity across outpatient departments. This model provides new channels for patients who previously could only choose between exiting the system or waiting for the next time slot when facing long queues or waiting times.

The optimization results provide scientific references for effectively improving consultation processes, with model concepts applicable to hierarchical diagnosis and treatment systems, two-way referrals between medical institutions, and emergency scheduling for urban public health events. Based on this research, several extensions merit further investigation: considering two-way overflow systems from smart healthcare or "Internet+" medical perspectives; employing reinforcement learning or clustering [24] algorithms for patient decision constraints; and incorporating more realistic overflow constraints including differentiated patient preferences, condition severity, and medical costs.

References

[1] Central People' s Government of the People' s Republic of China. The National Health Commission held a regular press conference on the quality and safety of medical services in China in 2019 [EB/OL]. (2020-10-17) [2021-09-27]. http://www.gov.cn/xinwen/2020-10/17/content_{5551912}.

[2] Du Yanping, Wu Jianceng. Practice and improvement of process optimization in hospital outpatient management [J]. Modern Hospitals, 2020, 20(06): 827-829.

[3] Joseph J W, Davis S R, Wilker E H, et al. Emergency physicians' active patient queues over the course of a shift [J]. American Journal of Emergency Medicine, 2020, 46(8): 254-259.

[4] Andersen A R, Nielsen B F, Reinhardt L B, et al. Staff optimization for time-dependent acute patient flow [J]. European Journal of Operational Research, 2019, 272(1): 94-105.

[5] Zhang A, Zhu X M, Lu Q, et al. Impact of prioritization on the outpatient queuing system in the emergency department with limited medical resources [J]. Symmetry Basel, 2019, 11(6): 796-797.

[6] Barros O, Weber R, Reveco C. Demand analysis and capacity management for hospital emergencies using advanced forecasting models and stochastic simulation [J]. Operations Research Perspectives, 2021. doi: https://doi.org/10.1016/j.orp.2021.100208.

[7] Wu X D, Li J, Chu C H. Modeling multi-stage healthcare systems with service interactions under blocking for bed allocation [J]. European Journal of Operational Research, 2019, 278(3): 927-941.

[8] Li Linyang, Li Chenyang. Optimization of hospital toll window setting based on queuing theory [J]. Journal of Hebei University: Natural Science Edition, 2020, 40(05): 449-453.

[9] Jing Tong, Ye Qingqing, Zhang Xiaoliang, et al. The performance analysis of a retrial M/M/1 queue with working breakdowns and balking [J]. Journal of Chongqing Normal University: Natural Science, 2020, 37(05): 1-9.

[10] Yang Ying, Li Jing. Study on patients' decision making modeling in specialist outpatient appointment based on queuing behavior [J]. Journal of Chongqing Normal University: Natural Science, 2019, 36(01): 21-28.

[11] He Q D, Huo J Z, Pan Y. The optimization model for the service process of stomatology department via DES simulation [J]. Cogent Business & Management, 2020, 7(1): 78-81.

[12] Kang D Y, Cho K J, Kwon O, et al. Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services [J]. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, 2020, 28(8): 325-331.

[13] Liu Jian, Zhao Hongkuan, Liu Sifeng. A study of service pricing with unfairness averse customers [J]. Chinese Journal of Management Science, 2018, 26(02): 46-53.

[14] Chen Ding, Fang Zhigeng, Liu Sifeng. Grey BD prediction model for spare parts of repairable queuing system [J]. Systems Engineering—Theory&Practice, 2020, 40(05): 1326-1338.

[15] Lu Yiling, Li Junxiang, Qi Yuan. Queuing model with variable service rate considering fatigue under epidemic situation [J/OL]. Journal of University of Shanghai for Science and Technology, 2021. (2021-09-30) [2022-02-23]. https://kns.cnki.net/kcms/detail/31.1739.t.20210929.1137.001.html

[16] Sun Jingxuan, Gao Jun, Zhang Junchuan, et al. M/M/C queuing model with the state of matching and abandoning [J]. Mathematics in Practice and Theory, 2020, 50(19): 186-192.

[17] Zhang Lifan, Jiao Pengpeng, Si Mingkai. Delay model of vehicles at urban road drop-off area [J]. Journal of System Simulation, 2020, 32(09): 1839-1846.

[18] Wang Jingwen, Li Huaqiang, Li Xuxiang, et al. Utility model and demand assessment method of integrated energy service [J]. Proceedings of the CSEE, 2020, 40(02): 411-425.

[19] Mariani F, Polinesi G, Recchioni M C. A tail-revisited Markowitz mean-variance approach and a portfolio network centrality [J]. Computational Management Science, 2022. DOI: 10.1007/S10287-022-00422-2.

[20] Xu Hongli, Liu Yuhao, Xu Wei, et al. Route choice model in stochastic network and its application in traffic assignment with consideration to travelers' decision inertia [J]. Systems Engineering—Theory&Practice, 2021, 41(04): 1010-1017.

[21] Yu Miao, Gong Jun, Tang Jiafu, et al. Decision model of delay information for call centers based on consideration of customers' patience [J]. Industrial Engineering and Management, 2015, 20(06): 108-113.

[22] Li Junxiang, Zang Wanbin. Simulation research on abandonment rate of contact center based on patience threshold [J]. Journal of System Simulation, 2021, 33(01): 169-179.

[23] Li Wuqiang, Liu Dezhi, Xu Xiaoqing. The effect of expert wages and service optimization on service policy when providing remedial service [J]. Chinese Journal of Management, 2020, 17(10): 1554-1563.

[24] Geng Xiuli, Wang Zhuxin. Recommendation of differential privacy scheme considering user interest analysis [J]. Application Research of Computers, 2022, 39(02): 474-478.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.