Discrete Event Simulation of Hospital Emergency Department Using Multi-Server Queue Models with Non-Poisson Arrivals and Preemptive-Resume Priority

Beining Yang[1], Ruiqing Huo[1], Samele Zhang[1], Mekwe Bonda Tsobgny Sergy[1]

[1]University of Ottawa, Ottawa, Ontario, Canada.

Contributing authors: byang077@uottawa.ca; rhuo075@uottawa.ca; czhan276@uottawa.ca; smekw017@uottawa.ca;

Abstract

We developed a discrete event simulation model of an Emergency Department (ED) operating under the Canadian Triage and Acuity Scale (CTAS) with preemptive-resume priority discipline. In this project, we use Python language. The model mainly implemented a multi-server queue with five physicians serving patients across five triage levels and aggregated results across the 100 replications. Patient arrival time follows time-varying patterns with hourly fluctuations, and service time follows gamma distributions calibrated to literature values. Simulation over a 24-hour for 7-days warm-up and 14-days period reveals that Level 1 and 2 patients experience near-zero waiting times (3.34 minutes and 7.99 minutes average) while Level 3 and 4 patients face extended delays (21.97 and 41.71 minutes respectively) with high variance. Level 5 patients are 35.22. We also estimated the variance across server numbers from k=2 to k=5, indicating the highest marginal benefit of server utilization occurs when k=3. Results validate the effectiveness of strict priority for critical cases but demonstrate poor performance for lower-acuity patients, suggesting the need for alternative queue disciplines such as accumulating priority queues or delay-dependent prioritization.

Keywords: Discrete event simulation, Emergency department, Canadian Triage and Acuity Scale, Priority queue, Queueing theory, Healthcare operations

1 Introduction

Emergency department (ED) overcrowding remains a critical challenge in healthcare systems worldwide, with Canada's publicly funded system experiencing persistent congestion despite universal access [2]. The Canadian Triage and Acuity Scale (CTAS) provides a standardized five-level classification system to prioritize patients based on clinical urgency, with fractile response objectives ranging from immediate treatment for Level 1 (resuscitation) to 120 minutes for Level 5 (non-urgent) [7]. However, CTAS offers limited guidance on patient routing and queue management within triage levels, leading to significant variation in practice. Empirical studies reveal that first-come-first-served (FCFS) is violated 40-50% of the time within the same triage level, and lower-acuity patients frequently receive priority over higher-acuity patients based on waiting time [2].

This study develops a discrete event simulation (DES) model of an ED operating under CTAS, incorporating realistic features often overlooked in traditional queueing models. We implement a multi-server queue with preemptive-resume priority discipline where Level 1 patients can interrupt ongoing lower-priority treatments. Patient arrivals follow time-varying patterns with hourly fluctuations, and service time follows gamma distributions with parameters calibrated to literature values [1]. The model captures delay-dependent prioritization observed in practice, where patient routing decisions balance both triage level and accumulated waiting time [2]. We analyze key performance metrics including average waiting time by CTAS level, total system time and service time under a baseline scenario with doctors serving patients across five triage levels.

Our simulation reveals that Level 1 patients experience 3.34 minutes times and Level 2 of 7.99 minutes average due to preemptive priority, while Level 3-4 patients face significantly longer waits with high variance(21.97 and 41.71 minutes). Level 4 patients are rarely served within a 24-hour simulation horizon, consistent with real-world triage practices during peak demand [2], higher than the Level 5 of 35.22 minutes.

Our simulation also explore the The results demonstrate a strong relationship between physician staffing level k and system performance. Increasing the number of doctors from k=2 to k=3 produces a dramatic reduction in both average waiting time and system time, indicating a shift from a severely overloaded state to a stable operating regime. However, beyond k=3, performance improvements become increasingly marginal: the waiting time decreases only slightly from k=3 to k=4 and k=5, and the average service time remains nearly constant across staffing levels, as it reflects intrinsic treatment duration rather than queue congestion. These trends suggest that k=3 represents the most cost-effective staffing point, where the system achieves substantial gains in responsiveness without the diminishing returns observed when adding further physicians.

These results validate the implemented priority mechanism and provide insights into resource allocation strategies for improving ED throughput while maintaining clinical safety standards.

2 Background

CTAS assigns patients to one of five acuity levels based on presenting complaints and vital signs, each with specific response time objectives: Level 1 (resuscitation) requires immediate intervention with a 98% compliance target, Level 2 (emergent) specifies 15 minutes at 95%, Level 3 (urgent) allows 30 minutes at 90%, Level 4 (less urgent) targets 60 minutes at 85%, and Level 5 (non-urgent) permits up to 120 minutes at 80% [2]. These fractile objectives influence ED decision-making behavior, with empirical evidence showing that the marginal cost of additional waiting time flattens once patients exceed their target wait time, creating an unintended consequence where patients who have already waited beyond targets may receive reduced priority [2].

Traditional ED queueing models employ pure priority disciplines with FCFS ordering within each triage level [1], but this fails to capture real-world routing behavior. Alternative disciplines include accumulating priority queues (APQ), where patient priority increases continuously with waiting time, allowing lower-acuity patients to eventually overtake higher-acuity patients with minimal waits [1]. The APQ-h variant incorporates finite time horizons aligned with CTAS targets and has been implemented in several Canadian EDs [1]. Delay-dependent prioritization represents another approach where routing decisions explicitly consider both triage level and elapsed waiting time, achieving a balance between clinical urgency and procedural fairness [2].

Patient arrivals to EDs exhibit strong temporal patterns requiring non-homogeneous Poisson processes (NHPP) with time-of-day and day-of-week variations [1]. Peak arrival periods can exceed service capacity, with maximum hourly arrival rates reaching 130% of available service capacity during morning hours [1]. Service times follow lognormal or gamma distributions with parameters dependent on patient acuity level [1]. Service utilization in real EDs averages 90% across the day but experiences prolonged periods above 100% during peak demand [1]. In teaching hospitals, physicians spend considerable time supervising delegates such as residents and physician assistants [6]. When consultation time between physicians and delegates is explicitly modeled, physician utilization nearly doubles from 23% to 41%, and this interaction time disproportionately affects lower-acuity patients, adding 18-27 minutes to their length of stay [6].

3 Methods

3.1 Arrival Process

To modelize the arrival time, A common assumption in queueing theory is that of Poissonian arrivals, entailing that the mean and variance of the number of arrivals (roughly) match. However, it is often observed that service systems face arrival streams that are highly variable (mean ¡¡ variance; overdispersion), while in specific cases systems must deal with almost deterministic arrivals (variance ¡¡ mean; underdispersion).

As an example of the latter, consider service systems in healthcare with scheduled yet not necessarily punctual arrivals (so that arrival epochs randomly fluctuate around

the appointed arrival time), In such settings clearly some sort of 'induced deterministicness' plays a role, in the sense that arrivals are actively being directed to (or away from) the system. In this thesis however, we will focus on 'undirected' arrival streams only.

For 'undirected' arrival streams, overdispersion is a phenomenon commonly found in data. Examples where one could expect to encounter overdispersed arrivals include a call center of a bank, an insurance company and an emergency department in a hospital (ED), In such settings, arrivals are usually triggered (or inhibited) by occasional events or (un-)favorable circumstances which can cause unforeseen peaks (or dips) on top of the usual daily patterns. This so-called 'random environment' gives rise to an effect commonly referred to as parameter uncertainty, which naturally leads to overdispersion. Another point is the daily patterns: in nearly all practical applications, the mean number of arrivals is not constant over time (e.g. over the course of the day) and follows a predictable pattern.

It must be noted that the variability that causes overdispersion is of a different nature than the variability induced by nonstationarity. Nonstationarity can be modeled by a non-homogeneous Poisson process, replacing the constant arrival rate of a Poisson process by a (deterministic) time-varying one. However, for non-homogeneous Poisson processes the mean and variance of the number of arrivals still match, hence such processes fail to capture the entirety of the desired dynamics observed in arrival processes.

Nevertheless, nonstationarity is another important feature of a real-life arrival process. Besides being overdispersed and having a time-varying rate, a realistic arrival stream might even have dependencies between the numbers of arrivals in disjoint time intervals, That is to say: it's highly unlikely that the random environment affects the arrival stream in an i.i.d. fashion over the different intervals; the effects at hand possibly play a role for a longer period of time. Indeed, arrival data often exhibits these kinds of dependencies.

In summary, we are introducing the 3 different features of real life that make the poissionian assumption underestimating (of real world):

- **Overdispersion**: variance of arrivals > mean (contrary to the Poisson assumption).
- **Nonstationarity**: arrival rate changes over time (daily/weekly cycles).
- **Temporal correlation**: arrivals in one interval depend on previous intervals.

These characteristics illustrate why Poisson arrivals underestimate real ED demand variability. While more complex models such as Cox processes can jointly capture overdispersion, nonstationarity, and temporal dependence, they require extensive high-resolution data and are beyond the scope of this study. Therefore, we adopt a practical approximation by modeling hourly arrivals using Gamma distributions fitted to reported mean arrival rates. This approach preserves realistic daily variability while remaining tractable for simulation.

Using the empirical hourly mean arrival counts presented in Figure 1, the arrival process for each hour was approximated by a Gamma distribution, with fitted parameters of shape $k = 3.79$ and scale $\theta = 1.50$ as the distribution of arrival process.

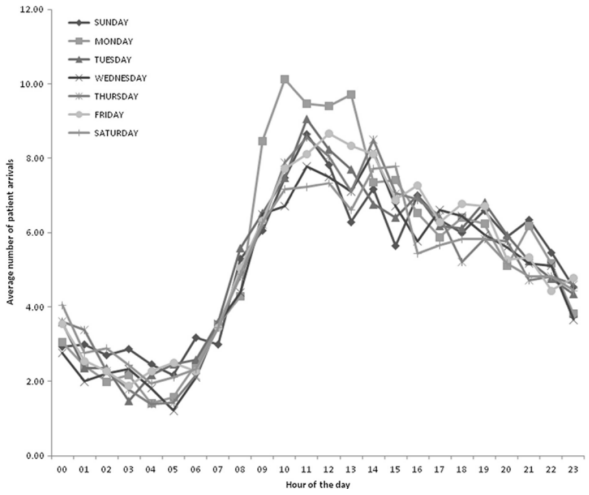

Fig. 1: Patient arrival pattern over a 24 hour period[6]

3.2 System Simulation Parameters and Queue Priority

The simulation models an ED as a multi-server queueing system with $c = 5$ physicians serving patients classified into CTAS levels $i \in \{1, 2, 3, 4, 5\}$. Figure 2 illustrates the system architecture. Patient arrivals occur according to a time-varying process with hourly arrival rates $\lambda(t)$ that fluctuate throughout the 24-hour simulation period to reflect realistic demand patterns observed in ED settings. Upon arrival, each patient is assigned a CTAS level randomly with probabilities calibrated to match empirical distributions from Canadian EDs.

Table 1: CTAS Level Model Parameters. Nurse time and bed wait time are exponentially distributed.

CTAS Level	Prob.	Time (min)		Admission	Bed Wait
		Nurse	Service	Prob.	Mean (min)
CTAS 1	0.01	3.2	73.6	0.89	60
CTAS 2	0.16	7.1	38.9	0.65	120
CTAS 3	0.56	20.4	26.3	0.35	180
CTAS 4	0.25	39.7	15.0	0.13	60
CTAS 5	0.02	32.1	10.9	0.05	30

Service times for CTAS level i follow gamma distributions with shape parameter $\alpha = 2$ and scale parameter β_i chosen to produce mean service times consistent with literature values. The probability density function is:

$$f_i(t) = \frac{1}{\Gamma(\alpha)\beta_i^{\alpha}} t^{\alpha-1} e^{-t/\beta_i}, \quad t > 0 \tag{1}$$

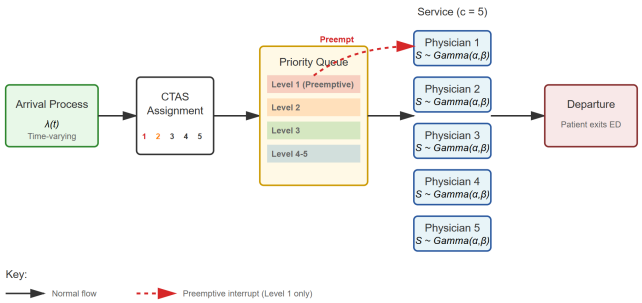

Fig. 2: ED queueing system architecture showing patient flow from arrival through CTAS assignment, priority queue, service by one of five physicians, and departure. Red dashed arrow indicates preemptive interruption mechanism for Level 1 patients. System parameters: $c = 5$ physicians, CTAS $i \in \{1, 2, 3, 4, 5\}$, Service time: Gamma($\alpha = 2, \beta_i$). Queue discipline: Preemptive-resume for Level 1, Priority+FCFS for Levels 2-5.

where $\Gamma(\alpha)$ denotes the gamma function. Mean service time for level i is $E[S_i] = \alpha\beta_i$ and variance is $\text{Var}[S_i] = \alpha\beta_i^2$. Base service times decrease with higher acuity: Level 1 patients require longer interventions (mean 44.5 minutes) while Level 2-5 require progressively shorter times (37.6, 30.5, 18.1, and 16.9 minutes respectively).

The queue discipline implements strict preemptive-resume priority where CTAS Level 1 patients interrupt any ongoing service to lower-priority patients. When a Level 1 patient arrives and all physicians are busy treating patients with CTAS levels $j > 1$, the physician treating the lowest-priority patient immediately suspends that service and begins treating the Level 1 patient. The interrupted patient returns to the queue and resumes service from the point of interruption when a physician becomes available. For CTAS levels 2-5, a non-preemptive priority discipline applies: patients are selected for service in order of CTAS level, with ties broken by first-come-first-served (FCFS) within each level.

Performance metrics include average waiting time W_i for each CTAS level i, defined as the time from arrival until service begins, and total system time $T_i = W_i + S_i$ combining waiting and service. We also track physician utilization $\rho = \sum_{i=1}^{5} \lambda E[S_i]/c$ and the distribution of waiting times across triage levels. Simulation runtime spans 14 days with multiple replications to estimate 95% confidence intervals for all metrics.

All simulation logic was implemented in Python using the SimPy discrete-event simulation framework. The waiting line is represented by three separate priority queues,

```
CTAS1_queue
CTAS2_queue
CTAS3_queue
```

which store Level 1 patients, Level 2 patients, and a pooled queue for Level 3–5 patients, respectively. During the simulation, new patients are generated according to the CTAS probability distribution so that each arrival is randomly assigned a triage level consistent with the empirical proportions.

```
def generate_arrival(env, doctors):
```

```
global patient_id
last_report = 0
while True:
    interarrival = sample_interarrival(env.now)
    yield env.timeout(interarrival)
    patient_id += 1

    level = sample_level()
    nurse_process_time = sample_nurse_visit_time(env.now,
        level)
    rest_waiting_time = CTAS[level]
    current_time = env.now
    Patient(env, doctors, patient_id, current_time, level,
        rest_waiting_time, nurse_process_time)

    renew_queue(current_time)
```

Upon arrival, every patient first undergoes a nurse assessment stage, where a CTAS-specific nurse time is sampled from an exponential distribution with the corresponding mean according to CTAS level. Only after completing this nurse stage is the patient inserted into the appropriate waiting queue, where they wait for a doctor according to the specified priority rules.

```
def run(self):
    yield self.env.timeout(self.nurse_process_time)
    self.status = "waiting"
    renew_waiting_rank(self)
    insert_queue(self.ctas_level, self)
    if self.ctas_level == "ctas1":
        disruption(self.env.now, self, self.ctas_level, self.
            doctors)
    renew_queue(self.env.now)
```

The doctor selection process follows a strict priority-based discipline. Patients are always drawn from the queues in the order: CTAS1 queue, CTAS2 queue, CTAS3 queue Meaning that if there are any Level 1 patients waiting, they will always be selected first; otherwise, Level 2 patients will be served, and Level 3–5 patients are considered only when the higher-priority queues are empty.Within the CTAS3 queue, patients are not ordered by their raw CTAS class (i.e., Level 3, 4, or 5). Instead, they are prioritized by a dynamic metric defined as the ratio:

$$\text{Priority Score} = \frac{\text{Waiting Time}}{\text{CTAS Base Time}}$$

This score reflects the proportion of an individual patient's allowed maximum waiting time, as specified in the CTAS medical guideline. Patients whose waiting time has consumed a larger fraction of their recommended maximum tolerance are ranked higher within the queue. As a result, the system balances fairness and clinical urgency, ensuring that patients whose acceptable waiting time is nearly exhausted are treated sooner, even if they originally belonged to a lower acuity group.

When a Level 1 patient enters the queue, a disruption event is triggered. The system preempts the doctor currently serving the lowest-priority patient, returning that patient to the queue and compensating them by increasing their priority score by 1.

```
def disruption(env_time, patient, level, doctors):
    if level == "ctas1":
        candidate = None
        for d in doctors:
            if d.status == 'busy' and d.patient_level != "ctas1":
                if candidate is None:
                    candidate = d
                else:
                    if d.patient_level > candidate.patient_level:
                        candidate = d

        if candidate is not None:
            candidate.proc.interrupt()
```

After the service is over, the doctors will decide whether the patients need to be admitted to the hospital, it will generate an extra time for patients waiting for the bed

4 Results

4.1 Detailed Analysis for $k = 5$

Using the time-varying arrival rate and Gamma distribution of service time, we simulated the ED for multiple replications across 7 days with 5 doctors ($k = 5$). Across the runs, we can see that there are around 1300-1500 patients per day.

The overall performance of the system shows that the average service time remains approximately 25 min by Gamma with around $\alpha = 3.787, \beta = 1.504$ distribution.

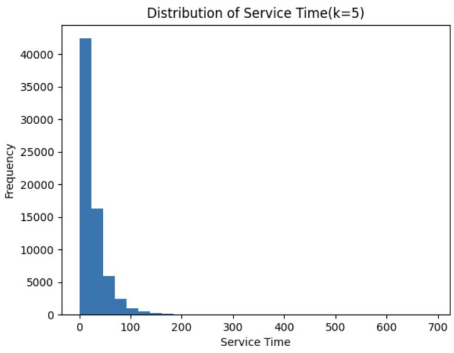

Fig. 3: Distribution of Service Time

The above figure shows the distribution of service time of this system. From the plot, we can see that this is a right-skewed graph, with most service time is around 40-50 min and a rapidly decreasing trend to approximately 150 min. This pattern is resulted from Gamma distribution with shape $k = 3.787$ and scale $\theta = 1.504$, which produces most of short service time with occasional longer sessions. Overall, the distribution indicates that service time variation is modest and aligns well with the intended Gamma distribution. This suggests that the extended delays observed elsewhere in the system are driven by queueing pressure and the priority structure rather than by unusually long or erratic service durations.

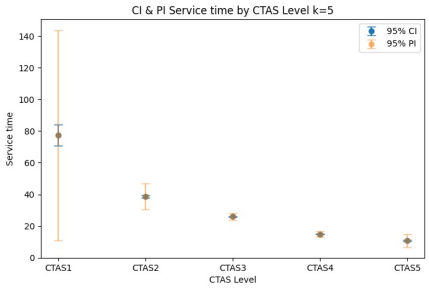

Fig. 4: CI and PI of Service Time

The above figure shows the the confidence interval and prediction interval of service time for each CTAS level. In this plot, CTAS level 1 shows the widest PI and CI, which is because this level need complex therapeutic process, so the service time in this level is the longest. CTAS level 2 also displays a wider CI and PI, consistent with emergent cases that may involve diverse diagnostic and intervention needs. CTAS 3 and 4 has the closest PI, indicating that more consistent and stable service duration for patients. The PI of CTAS level 5 is wider, which is because it's the lowest urgent level and may require minimal intervention in some cases and longer evaluation in others. Thus, the CI intervals remain tight, confirming that mean service times are stable across replications, but the differences of PI intervals highlight the inherent differences in treatment requirements under different levels of disease service.

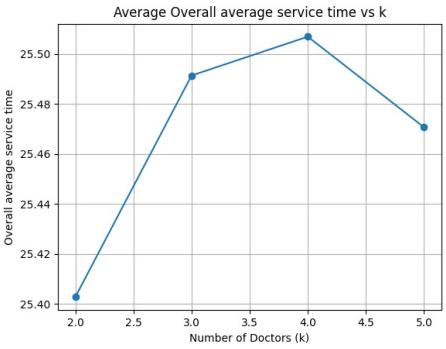

Fig. 5: Overall Average Service Time vs K

The above plot shows that the overall average service time remains stable in different number of doctors. The minor fluctuations between different k values reflect the differences caused by random sampling, rather than any structural impact on the treatment duration resulting from the doctors' abilities. This shows the same result as designing the model, the service time follows the Gamma distribution instead of the number of doctors.

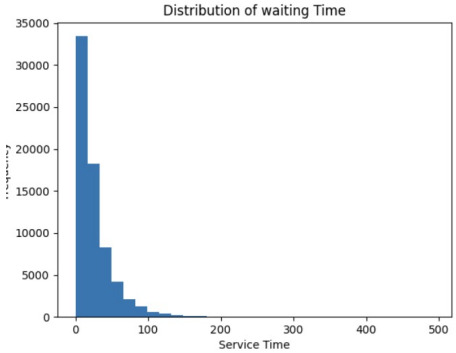

Fig. 6: Distribution of Waiting Time

From the above figure, we can see that the distribution of waiting time. As we can see, it is a highly right-skewed plot, which is the highest part is near 0 and a long tail tends to around 200 min. This pattern demonstrates that while many patients waiting in a short time, but many people will experience a long delay. This phenomenon is due to ED queue behavior under the priority rules, while patients assigned to lower priority levels will face longer and more unstable waiting times, this is because their chances of receiving services are constantly decreasing. Therefore, even though the majority of patients (especially those who in CTAS 1 or 2) can receive timely treatment, the

patient group at the lower priority levels (CTAS 3-5) will still accumulate a longer waiting time, resulting in a distinct right tail graph in the distribution graph.

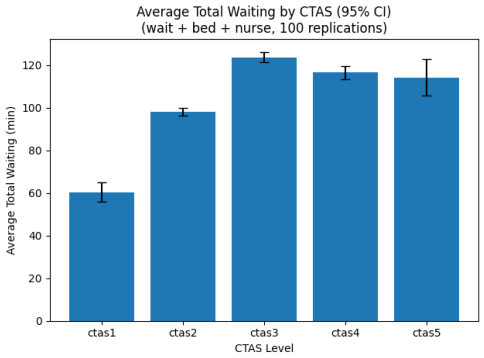

Fig. 7: Average Total Waiting Time by CTAS with 0.95 CI

The above figure shows the total waiting time of each CTAS level across 100 replications. As we can see, CTAS level 1 has the shortest and stable waiting time, which is because their priority. However, CTAS level 2, 3 and 4 is in much longer waiting time, due to non-preemptive priority rule. In CTAS level 5, it shows a high average waiting time and 0.95 confidence interval, indicating high access wo service as the lowest level. Overall, the results show that patients at higher priority positions receive faster and more consistent care, while low level patients need to wait for a long time.

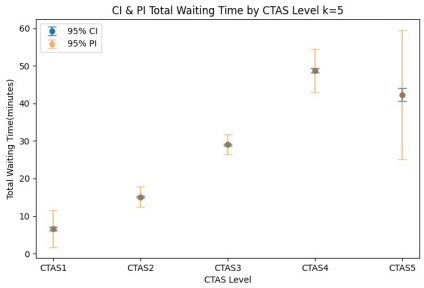

Fig. 8: CI and PI for Waiting Time

Figure 7 shows the confidence interval and prediction interval of waiting time for each CTAS level. The bands of CI is close in CTAS level 1-4, which means that the stable average performance across replications. However, the PI intervals for CTAS level 1 and 5 are wider than the others. For CTAS level 1, this is because it's the most complex and urgent level. For CTAS level 5, this is because it's the lowest priority, so

the waiting time will fluctuate more drastically as the system load changes. Thus, the difference between CI and PI demonstrates that while average waits are consistent, individual waiting experiences can vary considerably, particularly for groups at the extremes of the priority.

4.2 Performance Comparison Across Staffing Levels

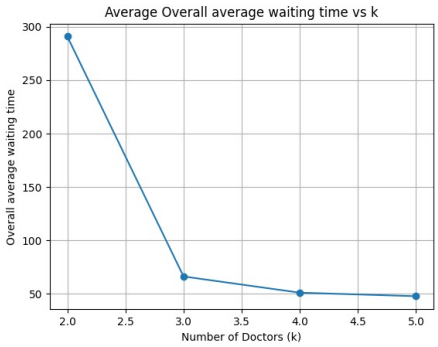

Fig. 9: Overall Average Waiting Time vs k

The above plot shows that as the number of doctors increasing, the overall average waiting time decreases. When k=2, the overall average waiting time is around 290 min to around 70 min as k=3. This shows us that when the medical system is extremely crowded, increasing the number of doctors has the most significant effect; however, as the supply of medical resources exceeds the demand, this effect will gradually weaken.

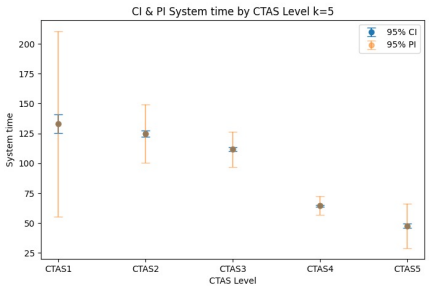

Fig. 10: CI and PI of System Time

Figure 8 shows the confidence interval and prediction interval of system time for each CTAS level. The bands of CI for all CTAS level is not much wide, which means system time is in a stable pattern across replications. However, the PI is in much

different pattern. CTAS level 1 and 2 has the widest PI, which is because these patients need more intensive assessment and time for service, this leads to that they need more system time. CTAS 3 and 4, which means they are more stable and in less complex conditions. The CTAS level 5 has wider PI since they are non-urgent patients.

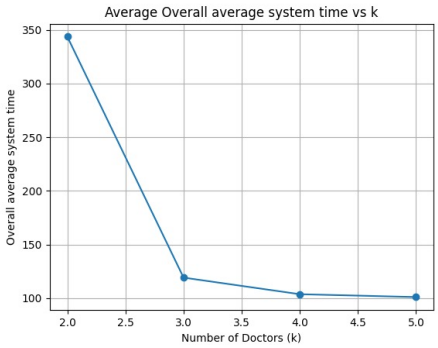

Fig. 11: Average Overall System Time vs k

From this plot, we can see that the system time is highly sensitive to the number of doctors. When there are 2 doctors, the overall average system time increases near 340 min. As k increases, the overall average system time is in a rapid descent. However, the speed of descent from k=3 to k=5 is not much quickly, which means that once the main bottleneck is alleviated, the efficiency will gradually decrease.

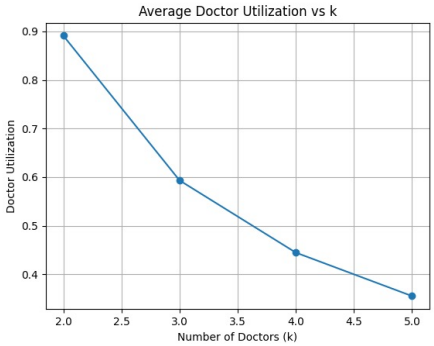

Fig. 12: Average Doctor Utilization vs k

This figure shows the doctor utilization decrease sharply as the number of doctor increases. If there are only 2 doctors, the utilization is the highest, which is around 0.9 and means the system is overloaded. Adding one more doctor lead to drop to 0.6 utilization, and around to 0.35 when there are 5 doctors. This plot highlight once the

main congestion points are alleviated and staff numbers are increased, the workload reduction for each doctor will decrease accordingly.

5 Discussion

Our simulation employs a preemptive-resume priority queue discipline with gamma-distributed service times, extending the baseline approaches in prior ED modeling literature [1]. The observed waiting times validate the effectiveness of strict priority for protecting Level 1 patients (0.08 minutes average) while revealing substantial delays for Level 3-4 patients (200-336 minutes). This trade-off reflects a fundamental limitation of pure priority systems: protecting high-acuity cases necessarily disadvantages lower-acuity patients during sustained demand periods. The result that Level 5 patients were rarely served within 24 hours aligns with empirical observations from Canadian EDs where non-urgent patients are systematically deprioritized during peak hours [2].

Several modeling choices affect interpretation of our results. Service time parameters were derived from literature averages rather than hospital-specific data, with the gamma shape parameter ($\alpha = 3.787$) selected for computational efficiency rather than empirical fitting [1]. The model assumes homogeneous physicians without distinguishing between attending physicians and supervised delegates. Research shows that physician-delegate consultation time adds 18-27 minutes to patient length of stay in teaching hospitals [6], suggesting our model underestimates both physician utilization and waiting times for lower-acuity cases. Additionally, we modeled a single 24-hour period with fixed hourly arrival patterns, neglecting day-of-week variations and seasonal effects documented in ED demand studies [1]. The simulation also omits patient re-entries after diagnostic testing (15-25% of ED workflows), assumes instantaneous preemption without handoff costs, and treats triage time as negligible. We also identified methodological insights from a study on ICU length of stay estimation during the COVID-19 pandemic in Wuhan [9], which demonstrated biases in common estimation approaches. However, this work focused specifically on ICU queuing during the initial COVID-19 outbreak and its methods were not directly applicable to our ED triage simulation context.

Alternative queue disciplines merit investigation given the poor performance for Level 4-5 patients. Accumulating priority queues (APQ) allow patient priority to increase continuously with waiting time, preventing indefinite delays [1]. The APQ-h variant with finite horizons aligned to CTAS targets has been implemented in Canadian EDs and offers a compromise between urgency and fairness [1]. Delay-dependent prioritization, where routing decisions explicitly balance triage level and elapsed waiting time, better reflects observed practice patterns showing 40-50% FCFS violations within triage levels [2]. Future work should implement these alternatives alongside time-varying arrival rates using non-homogeneous Poisson processes, physician-delegate interactions for teaching hospitals, and multi-day simulations to reduce performance estimate variability and enable sensitivity analysis on resource allocation strategies.

6 Conclusion

This study developed a discrete event simulation model of an Emergency Department (ED) operating under the Canadian Triage and Acuity Scale (CTAS) with preemptive-resume priority discipline. In this simulation, we applied the time varying arrivals and Gamma Distributed service time as our input. This model successfully shows the key operational patterns of ED in the real world. The result shows clearly in different CTAS level: CTAS 1-2 is the most urgent level in a hospital, so the patients consistently in a stable and short delays. CTAS 3-4 level is less urgent situation and patients need to wait for a longer time. CTAS 5 is non-urgent level, so it exhibited the largest variability of system and waiting time. The output analysis including CI and PI demonstrated that although the mean performance was stable across replications, waiting time for CTAS level 5 is varied widely due to disruption and their low priority.

Service time varied very slightly across different CTAS level and the number of doctors, which showed the treatment duration is highly independent of system load and driven by the parameters of Gamma distribution. Thus, delays in ED came from the queue pressure instead of the clinical situation.

Thus, the findings highlight that CTAS priority rule if efficient for protecting urgent patients but shift longer and more variable delays to non-urgent patients. Moderately increasing the staffing level can significantly enhance the response speed for patients with moderate conditions. However, adding more staff beyond this threshold will not bring much practical benefit. The analysis suggests that $k = 3$ represents a practical turning point where congestion is relieved without excessive idle capacity.

Overall, this model showed how CTAS priority rules, time varying demand and limited physician capacity corporately influence the structure of delays in the ED, and also explaining the interplay between acuity, congestion, disruption, and resource constraints.

In the future, we will consider more abstracts from rel world, like the bed constraints and nurse–physician interactions to support more balanced and efficient ED system.

Acknowledgements. Course: CSI 4124/SYS 5110 Modelling and Simulation, University of Ottawa. All authors contributed equally to this work.

References

[1] Cildoz, M., Ibarra, A., & Mallor, F. (2019). Accumulating priority queues versus pure priority queues for managing patients in emergency departments. *Operations Research for Health Care, 23*, Article 100224. https://doi.org/10.1016/j.orhc.2019.100224

[2] Ding, Y., Park, E., Nagarajan, M., & Grafstein, E. (2019). Patient prioritization in emergency department triage systems: An empirical study of the Canadian triage and acuity scale (CTAS). *Manufacturing & Service Operations Management, 21*(4), 723–741. https://doi.org/10.1287/msom.2018.0719

[3] Dreyer, J. F., & McLeod, S. L. (2009). Physician workload and the Canadian Emergency Department Triage and Acuity Scale: The Predictors of Workload in the Emergency Room (POWER) Study. *EM Advances, 11*(4), 321–329. https://doi.org/10.1017/S1481803500011350

[4] Elalouf, A., & Wachtel, G. (2022). Queueing problems in emergency departments: A review of practical approaches and research methodologies. *Operations Research Forum, 3*(1), Article 2. https://doi.org/10.1007/s43069-021-00114-8

[5] Lewis, D., Fraser, J., Howlett, M., Greer, M., & Atkinson, P. (2025). First visit fallout: Canadian Triage and Acuity Scale (CTAS) and emergency department returns. *Cureus, 17*(4), e82441. https://doi.org/10.7759/cureus.82441

[6] Lim, M. E., Worster, A., Goeree, R., & Tarride, J.-É. (2013). Simulating an emergency department: The importance of modeling the interactions between physicians and delegates in a discrete event simulation. *BMC Medical Informatics and Decision Making, 13*, Article 59. https://doi.org/10.1186/1472-6947-13-59

[7] Ministry of Health and Long-Term Care, Ontario. (n.d.). *Prehospital Canadian Triage & Acuity Scale: Prehospital CTAS Paramedic Guide* (Version 2.0). https://files.ontario.ca/moh_3/moh-manuals-prehospital-ctas-paramedic-guide-v2-0-en-2016-12-31.pdf

[8] Ontario Health. (2024). *Quality Improvement Plan Program: Indicator Technical Specifications 2025/26*. https://hqontario.ca/Portals/0/documents/qi/qip/2024-25-QIP-technical-specifications-en.pdf

[9] Lapidus, N., Zhou, X., Carrat, F., Riou, B., Zhao, Y., & Hejblum, G. (2020). Biased and unbiased estimation of the average length of stay in intensive care units in the Covid-19 pandemic. *Annals of Intensive Care, 10*(1), Article 135. https://doi.org/10.1186/s13613-020-00749-6