Saudi Journal of Emergency Medicine

REVIEW ARTICLE

Emergency department crowding through the lens of queuing theory

Fandi Z. Alanazi[1], Jonathan A. Edlow[2], David T. Chiu[3]

ABSTRACT

Emergency department crowding remains a persistent challenge across health systems worldwide and is associated with delayed care, increased risk of adverse events, and clinician burnout. Conventional explanations often focus on rising patient volumes or local operational inefficiencies, yet these factors do not fully explain why crowding frequently occurs even in well-resourced hospitals. Queuing theory, a framework from operations research, provides a useful lens for understanding how demand, capacity, and variability interact in complex service systems. When applied to emergency departments, queuing theory demonstrated how high utilization, variability in arrivals and service times, and downstream constraints such as inpatient bed availability can produce sudden congestion and prolonged delays. Viewing emergency departments as networks of interconnected queues helped explain why localized interventions often failed to improve flow and why system-level coordination is required. Although this article focused on emergency department operations, these principles applied broadly across hospital environments where variable demand interacts with constrained resources. Incorporating queuing theory into health system planning might provide clinicians and administrators with a clearer framework for designing more resilient systems capable of delivering timely care under conditions of uncertainty.

Keywords: Emergency department crowding, queuing theory, patient flow, hospital operations, health systems, review article.

Introduction

Emergency department crowding continues to pose a significant challenge across healthcare systems worldwide [1]. Despite substantial investments in infrastructure, staffing, and technology, many emergency departments continue to experience prolonged waiting times, access restrictions, and frequent periods of operational stress. Clinicians working in these environments often perceive crowding as a consequence of rising patient volumes, staffing limitations, or episodic surges in demand. While these factors contribute, they do not fully explain why crowding frequently occurs even in well-resourced departments staffed by experienced clinicians.

This apparent paradox has important implications. Emergency department crowding is associated with delays in care, increased risk of adverse events, reduced patient satisfaction, and significant moral distress among healthcare workers [2]. Traditional responses often focus on working faster, adding staff, or expanding physical space. However, such interventions frequently provide only temporary relief and might fail to produce sustained improvements in flow.

Queuing theory offers a complementary and system-level explanation for this phenomenon. Originating from operations research, queuing theory provides a structured framework for understanding how demand, capacity, and variability interact in complex service systems. When applied to emergency medicine, it explains why small changes in patient arrivals or service times can lead to disproportionate increases in waiting and congestion. Although this article focuses on patient flow within the emergency department, the underlying principles of queuing theory apply broadly to many departments and operational environments throughout the hospital. Importantly, queuing theory demonstrated that crowding is not necessarily a failure of effort or commitment but rather a predictable outcome of how systems behave under high utilization and variability.

Correspondence to: Fandi Z Alanazi
Department of Emergency Medicine, Imam Abdulrahman Bin Faisal University, King Fahad University Hospital, Dammam, Saudi Arabia.
Email: fandizaed@gmail.com
Full list of author information is available at the end of the article.
Received: 03 June 2026 | **Accepted**: 28 July 2026

OPEN ACCESS
This is an open access article distributed in accordance with the Creative Commons Attribution (CC BY 4.0) license: https://creativecommons.org/licenses/by/4.0/) which permits any use, Share — copy and redistribute the material in any medium or format, Adapt — remix, transform, and build upon the material for any purpose, as long as the authors and the original source are properly cited. © The Author(s) 2025.

The review aimed to introduce the core concepts of queuing theory in accessible, non-mathematical language and to demonstrate how these principles apply directly to emergency department operations. By translating abstract theory into clinically meaningful insights, this article aimed to equip emergency physicians and healthcare leaders with a deeper understanding of crowding and flow and to support more effective, system-oriented solutions within emergency care settings around the world.

Results and Discussion

What is queuing theory?

Queuing theory is the study of systems in which entities arrive over time, wait for limited resources, receive service, and then depart under defined constraints. Rather than focusing on individual performance, it examines how entire systems behave when demand approaches or exceeds available capacity [3]. In emergency medicine, queuing models provide simplified representations of how patients enter the emergency department, how care is delivered using finite clinical and physical resources, and how delays and congestion emerge. These models do not attempt to reproduce clinical care in detail. Rather, they capture the fundamental dynamics of patient flow, waiting, and system instability in high-variability environments [4].

At its most basic level, an emergency department queuing model consisted of five elements. First, arrivals described how patients enter the system. In emergency departments, arrivals are largely unscheduled and random, occurring through walk-ins or ambulance arrivals and varying seasonally but also by time of day and day of the week. Furthermore, each arrival is of varying complexity and resource needs.

Second, servers represented who or what provides care. In the emergency department, servers include physicians, nurses, treatment beds, resuscitation bays, and diagnostic resources such as laboratory and radiology services.

Third, service time reflected how long care takes. This might include the time from initial physician assessment to a disposition decision and is highly variable depending on patient acuity, diagnostic complexity, resource utilization, and consultation requirements.

Fourth, queue rules determine who waits and who is served next. Unlike many service systems, emergency departments use priority-based queuing, where higher-acuity patients are seen first according to triage systems such as the Emergency Severity Index rather than simple first-come, first-served order.

Finally, exit constraints described how patients leave the system. In emergency medicine, exits included discharge home, observation, or inpatient admission. When inpatient beds are unavailable, admitted patients remain in the emergency department, creating boarding delays that effectively reduce system capacity.

Together, these elements already constitute a complete queuing model for emergency department flow [5]. Even this simplest representation is sufficient to explain, many of the operational challenges observed in emergency departments. When arrival rates increase, service times lengthen, or exit constraints tighten, queues grow and waiting times rise, often rapidly and unpredictably [6,7].

The tipping point in emergency departments

A central concept in queuing theory is utilization, defined as the proportion of available system capacity that is in use at a given time. In emergency departments, utilization reflected the relationship between patient arrival rates and the capacity of clinicians, treatment spaces, diagnostic services, and downstream inpatient beds to deliver care.

High utilization is often perceived as a marker of efficiency in many service industries because unused capacity might appear wasteful from a financial or resource-allocation perspective.

In hospitals, for example, empty beds or idle treatment spaces can be interpreted as underutilized resources. From a queuing perspective, however, sustained high utilization represented system fragility rather than efficiency.

When most available capacity is already in use, little reserve remains to absorb normal variability in patient arrivals, service times, or downstream delays, making the system highly vulnerable to congestion.

Queuing theory demonstrated that waiting time does not increase linearly as utilization rises. At moderate levels of utilization, emergency departments can absorb normal fluctuations in demand with minimal impact on flow. Empirically and theoretically, systems operating below approximately 80% utilization are generally able to recover from short-term surges without prolonged delays. In this range, performance appears stable, and waiting times remain relatively predictable [8].

As utilization approached approximately 85%, the system enters a fragile operating zone. In this range, small increases in patient arrivals, minor delays in diagnostics, or brief reductions in staffing can produce disproportionately large increases in waiting time. The emergency department might still appear functional, but its ability to absorb variability is markedly reduced. Recovery from transient surges becomes slower and less reliable.

Beyond approximately 90% utilization, queuing systems become operationally unstable [9,10]. Waiting times escalate rapidly, queues grow faster than they can be cleared, and crowding becomes persistent rather than episodic. At this level of utilization, even small disruptions can push the system into prolonged congestion. In addition, prolonged waiting itself generates additional work for the ED [11]. As patients remain in the ED, they require repeated reassessments, additional monitoring, medication adjustments, nursing care, and coordination tasks that would not occur in a system with timely throughput. This secondary workload effectively lengthens service times, further increasing utilization and accelerating congestion.

When utilization approaches 95% or higher, the system enters a collapse zone, characterized by near-continuous

crowding, delayed care, and difficulty returning to baseline operations without external intervention.

Emergency departments frequently operate near or within these high-utilization zones. Unlike scheduled care environments, emergency medicine must accommodate unpredictable arrivals, wide variation in patient acuity, and frequent interruptions in workflow. When baseline utilization is already high, there is little reserve capacity available to buffer these unavoidable sources of variability [7].

Variability is a critical amplifier of these effects. Variability exists in arrival volumes, arrival patterns, service times, and exit capacity. Even when average demand appears stable, emergency department arrival volumes fluctuate around a mean [7]. Some days would naturally exceed the average by substantial margins, and these high-volume periods can rapidly push the system toward its capacity limits. Smaller emergency departments are particularly vulnerable to these fluctuations because even modest increases in arrivals can have a disproportionate impact on system capacity.

Variability also occurs in arrival patterns within a given day. Clustering of ambulance arrivals, prolonged diagnostic evaluations, complex consultations, and inpatient boarding all increase variability. Queuing theory showed that variability magnifies congestion most severely when utilization is high [10,11]. Two emergency departments with similar average volumes and staffing might therefore experience markedly different levels of crowding depending on how variability is distributed and managed.

Exit constraints further accelerated the approach to the tipping point. When admitted patients remain in the emergency department due to limited inpatient bed availability, they continue to occupy treatment spaces and consume clinical attention. From a queuing perspective, this represented a restriction at the exit of the system, effectively reducing service capacity and increasing utilization [6,11]. As exit capacity tightens, the emergency department is pushed closer to instability even if arrival volumes remain unchanged [12].

These dynamics explained a common clinical experience. Emergency departments often appear to function adequately for extended periods, followed by sudden and severe crowding without a clear triggering event. Queuing theory predicted this behavior. Systems operating near capacity can appear stable until a threshold is crossed, after which waiting times increase rapidly, and recovery becomes difficult.

Understanding utilization and its interaction with variability reframed how emergency department crowding is interpreted. Crowding is not primarily the result of inadequate effort or inefficient clinicians; rather, it is the predictable consequence of operating a high-variability system near its capacity limits. Sustainable solutions therefore require not only sufficient resources, but deliberate protection of reserve capacity and system-level strategies to reduce unnecessary variability across the continuum of care.

Emergency departments as networks of queues

Emergency departments are often described as single crowded spaces, but from a queuing perspective they function as networks of interconnected queues. Patients do not wait in only one line. Instead, they move through a sequence of queues, each with its own demand, capacity, and sources of delay [13]. Understanding EDs as networks rather than isolated workstations is essential for explaining why crowding persisted and why interventions targeting a single bottleneck often failed to produce sustained improvements.

The patient journey typically begins with triage, followed by initial clinical assessment, diagnostic testing, consultation, treatment, and disposition. Each of these stages can be viewed as a distinct but interconnected queue with its own arrival patterns, service times, and resource constraints [4,11]. Delays at any stage propagate both downstream and upstream, affecting the performance of the entire system. A seemingly minor delay in diagnostic imaging or specialty consultation can quickly translate into prolonged bed occupancy and increased waiting for newly arriving patients.

For example, once a patient enters the queue for a CT scan, the diagnostic result depends on passage through several additional queues before it can inform clinical decision-making. These might include a queue for intravenous access, a queue for patient transport to the scanner, the queue for the CT scanner itself, transport back to the emergency department, and finally the queue for radiologist interpretation. Delays at any one of these queues prolong diagnostic decision time, increase patient length of stay, and reduce the availability of treatment spaces for newly arriving patients.

Importantly, improving performance at one stage of the network might worsen congestion elsewhere. For example, faster triage and initial physician assessment can increase the rate at which patients enter queues for imaging, laboratory testing, or inpatient admission. If capacity in those downstream processes does not increase simultaneously, congestion simply shifts location rather than resolving [14]. From a queuing perspective, this phenomenon reflected the interdependence of queues within a networked system

Emergency department crowding is therefore not solely determined by front-end processes. Output constraints play a dominant role. Admission to inpatient units represented the final queue in the emergency department network. When inpatient beds are unavailable, admitted patients remain in the emergency department, occupying treatment spaces and consuming nursing and physician attention. This boarding phenomenon effectively converts the emergency department into an extension of the inpatient ward, reducing effective treatment capacity while simultaneously diverting clinical attention from newly arriving patients, thereby amplifying workload and worsening congestion across the entire network.

These dynamics helped explain why emergency department-based interventions alone often failed to resolve crowding. Adding staff, opening fast-track areas, or optimizing triage processes might improve

performance at specific nodes but would not overcome downstream constraints. In networked queuing systems, the slowest or most constrained segment governs overall throughput. Sustainable improvements therefore require alignment across the emergency department and the broader hospital system [15].

Viewing emergency departments as networks of queues also highlighted the importance of coordination and timing. Variability introduced at one node can destabilize others, particularly when utilization is high. For example, elective surgical scheduling that concentrates admissions during specific hours might overwhelm inpatient capacity, increasing boarding and degrading emergency department flow. Conversely, smoothing demand and service across the network (e.g., scheduling elective surgery on the weekend) can improve performance without increasing total capacity [10].

From this perspective, emergency department crowding emerged as a system-level property rather than a localized operational failure. Queuing theory reframed crowding as the predictable result of interactions among multiple constrained processes operating under uncertainty. Recognizing emergency departments as networks of queues encourages solutions that extend beyond the physical boundaries of the department and engage hospital-wide leadership, bed management, and policy planning.

Conclusion

Emergency department crowding is often experienced as a daily operational frustration, yet queuing theory demonstrated that it is a predictable consequence of how complex, high-variability systems behave when operated near capacity. By framing emergency departments as queuing systems and networks of interconnected queues, this perspective explained why crowding can persist despite adequate staffing, strong clinical effort, and well-intentioned local interventions. A central insight from queuing theory is that crowding is not primarily determined by average patient volumes. Instead, it emerges from the interaction between utilization and variability.

Emergency departments might appear adequately resourced when judged by average daily census or mean length of stay, yet still experienced severe congestion when variability in arrivals, service times, or exit capacity coincided with high baseline utilization. Under these conditions, small disruptions can produce disproportionately large increases in waiting and delay. Understanding this distinction has important implications for emergency physicians and healthcare leaders. Sustainable solutions to crowding require moving beyond strategies that focus solely on managing average demand toward system-level approaches that protect reserve capacity and reduce unnecessary variability across the continuum of care. Incorporating queuing theory into emergency care planning provides a structured framework for designing more resilient emergency departments capable of delivering timely, safe, and reliable care under conditions of uncertainty.

List of abbreviations

CT- Computed Tomography

ED- Emergency Department

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

None.

Consent for publication

Not Applicable.

Ethical Approval

Not applicable.

Author details

Fandi Z. Alanazi[1], Jonathan A. Edlow[2], David T. Chiu[3]

1. Department of Emergency Medicine, Imam Abdulrahman Bin Faisal University, King Fahad University Hospital, Dammam, Saudi Arabia
2. Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
3. Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

References

1. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA. A conceptual model of emergency department crowding. Ann Emerg Med. 2003;42(2):173–80. https://doi.org/10.1067/mem.2003.302
2. Hoot NR, Aronsky D. Systematic review of emergency department crowding: causes, effects, and solutions. **Ann Emerg Med. 2008;52(2):126–36.e1. https://doi.org/10.1016/j.annemergmed.2008.03.014**
3. **Lakshmi C, Iyer SA.** Application of queueing theory in health care: a literature review. Oper Res Health Care. 2013;2(1–2):25–39. https://doi.org/10.1016/j.orhc.2013.03.002
4. Hu X, Barnes S, Golden B. Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies. Int Trans Oper Res. 2018;25(1):7–49. https://doi.org/10.1111/itor.12400
5. Joseph JW. Queuing theory and modeling emergency department resource utilization. Emerg Med Clin North Am. 2020;38(3):563–72. https://doi.org/10.1016/j.emc.2020.04.006
6. Forster AJ, Stiell I, Wells G, Lee AJ, van Walraven C. The effect of hospital occupancy on emergency department length of stay and patient disposition. Acad Emerg Med. 2003;10(2):127–33. https://doi.org/10.1197/aemj.10.2.127
7. Wiler JL, Bolandifar E, Griffey RT, Poirier RF, Olsen T. An emergency department patient flow model based on queueing theory principles. Acad Emerg Med. 2013;20(9):939–46. https://doi.org/10.1111/acem.12215
8. Petrie DA, Comber S. Emergency department access and flow: complex systems need complex approaches. J Eval

Clin Pract. 2020;26(5):1552–8. https://doi.org/10.1111/jep.13418

9. Bagust A, Place M, Posnett JW. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. BMJ. 1999;319(7203):155–8. https://doi.org/10.1136/bmj.319.7203.155

10. McManus ML, Long MC, Cooper A, Litvak E. Queuing theory accurately models the need for critical care resources. Anesthesiology. 2004;100(5):1271–6. https://doi.org/10.1097/00000542-200405000-00032

11. Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB. On patient flow in hospitals: a data-based queueing-science perspective. Stoch Syst. 2015;5(1):146–94. https://doi.org/10.1287/14-SSY153

12. Lin D, Patrick J, Labeau F. Estimating the waiting time of multi-priority emergency patients with downstream blocking. Health Care Manage Sci. 2014;17(1):88–99. https://doi.org/10.1007/s10729-013-9241-3

13. Saghafian S, Austin G, Traub SJ. Operations research/management contributions to emergency department patient flow optimization: review and research prospects. IIE Trans Healthc Syst Eng. 2015;5(2):101–23. https://doi.org/10.1080/19488300.2015.1017676

14. Ortiz-Barrios MA, Alfaro-Saíz JJ. Methodological approaches to support process improvement in emergency departments: a systematic review. Int J Environ Res Public Health. 2020;17(8):2664. https://doi.org/10.3390/ijerph17082664

15. Elalouf A, Wachtel G. Queueing problems in emergency departments: a review of practical approaches and research methodologies. Oper Res Forum. 2021;3(1):2. https://doi.org/10.1007/s43069-021-00114-8