Queuing Theory and Modeling Emergency Department Resource Utilization

Joshua W. Joseph, MD, MS, MBE

KEYWORDS

• Emergency department operations • Throughput • Queueing • Efficiency

KEY POINTS

- Many different processes within the emergency department (ED) can be modeled using queueing equations, which can predict delays and help to identify bottlenecks to throughput.
- Average arrival times often hide variability that can significantly increase waiting times.
- As a resource approaches a utilization ratio of 1, waiting times for the resource grow dramatically.
- Although queueing equations can become mathematically complex, software packages that simulate queues help to predict the effects of changing resources in the ED without the need for tricky calculations.

INTRODUCTION

How many doctors does it take to manage the waiting room? It sounds like a simple question, which should have a simple answer, solved with algebra. An emergency physician (EP) can see X patients per hour. Y patients per hour arrive in the emergency department (ED). Should not the number of EPs needed to keep the waiting room clear be the number of patients who arrive per hour, divided by the number of patients per hour an EP can see?

The more one thinks about the problem, however, the more complex it becomes, and the less satisfactory the simplifying assumptions. Patients take different amounts of time to evaluate; a healthy 20 year old with a sore throat does not take the same amount of time as a 65 year old on dialysis presenting with fever and chest pain. Much also depends on the resources available to the ED, and the ED's current level of activity. If the ED has to share a computed tomography (CT) scan with the rest of the hospital, or if there is already a glut of samples at the laboratory, the EP

Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, One Deaconess Road, Boston, MA 02215, USA
E-mail address: jwjoseph@bidmc.harvard.edu

Emerg Med Clin N Am 38 (2020) 563–572
https://doi.org/10.1016/j.emc.2020.04.006 emed.theclinics.com

can work flawlessly, but the number of patients in the waiting room will continue to grow.

This problem may have too many inputs and unknowns to be solved exactly, and variations within the workflow of the ED might lead two EDs with similar resources and patient volumes to different solutions. Yet, the question is too important to simply ignore or guess at. Staffing an ED has serious implications for its financial viability and for the safety and quality of care it delivers to its patients. It is also a question that has profound implications for the potential stresses and work-life balance issues that EPs and nurses may face, because the need to provide coverage for the ED at all times of day, every day of the year, imposes demands on their family life and circadian rhythms.

Thankfully, although we might not be able to create a perfect or exact answer to our question, we can make some important estimations. In the process, we can find out how to avoid several significant pitfalls that are commonly made by administrators when allocating resources in the ED. As the statistician George Box observed, "all models are wrong, but some are useful."[1]

MODELING RESOURCE USE AND STAFFING

Averages Are Not Enough

A simpler, related question to how many EPs are needed to staff an ED is how many nurses are needed to staff triage at a particular time. For most EDs, triage is a discrete process that takes a fixed amount of time. That the process is discrete means that we measure it in terms of individual patients, but for the purpose of modeling, we can also use discrete in the sense that triaging a patient is a self-contained process. The full patient evaluation conducted by an EP often contains many contingent stages, such as waiting for laboratory tests or imaging. Conversely, the typical steps of triage, such as identifying the patient, asking their chief complaint, obtaining vital signs, conducting a brief examination, and assigning a priority score (eg, ESI, ATS, or CTAS), are performed sequentially before seeing the next patient, and can be modeled as a single time period.[2–4]

A commonsense approach is to compare the average number of patients who arrive in the ED in a given period of time with the average number of patients that a nurse can triage in the same period of time. This makes some assumptions about the human factors involved, because we might be inclined to underestimate how quickly nurses can triage patients if some nurses can work faster under pressure for a period of time. However, obtaining a representative average that takes into account busy and calmer periods should help to even this out. Using this approach, an ED that averages 12 patients arriving per hour, where nurses can triage six patients per hour, would need two nurses to staff triage at a given time.

According to this approach, the two nurses intrepidly staffing the triage station should be ready to manage 12 patients per hour. Better yet, when the stars align and exactly 12 patients per hour arrive, our model suggests that as each patient arrives, one of the nurses will have just finished triaging the last arrival. Who would have thought that creating the "no wait" ED could be so simple.

Anyone who has worked in an ED knows that the results of this model do not match reality. However, it is difficult to explain exactly why this happens. It is tempting to accuse the averages of the time to triage a patient and the number of patients per hour of being misleading. Surely, our model hits a snag when the average underestimates the number of patients arriving per hour, and the waiting time builds up because during some hours the number of patients is higher than average, and perhaps there are times when our nurses are going more slowly.

Unfortunately, the averages are not really the problem; they remain an essential part of creating a representative model of an ED. Instead, the problem is in how we use the averages. We can make things even simpler, by imagining that in this ED, the average number of patients is ironclad, so that there are always 12 patients arriving per hour, and that the nurses always take 6 minutes to evaluate a patient. However, even in this imaginary ED, with triage staffed with two nurses, there will be times at which patients wait. Rarely, they may wait for a long time to be triaged. Even if our nurses become faster, or if we add a third nurse, it is still likely that some of the patients will have to wait.

The Problem of Arrivals and Variation

Waits arise in our ED because of variations in the amount of time between when patients arrive. If patients arrive at regularly spaced intervals (Fig. 1A), then they are triaged exactly as they arrive, and the arrivals perfectly match with the completion of the last patient's triage. However, variation in the amount of time between arrivals (Fig. 1B) means that there are idle periods in which one or more of the nurses does not have a new patient to triage, and unavoidable waiting periods in which a patient arrives while multiple nurses are already busy conducting a triage assessment.

Unlike when taking the average number of arrivals over a period of time, these losses of time do not even out. When a nurse is idle in triage and has no patient to see,

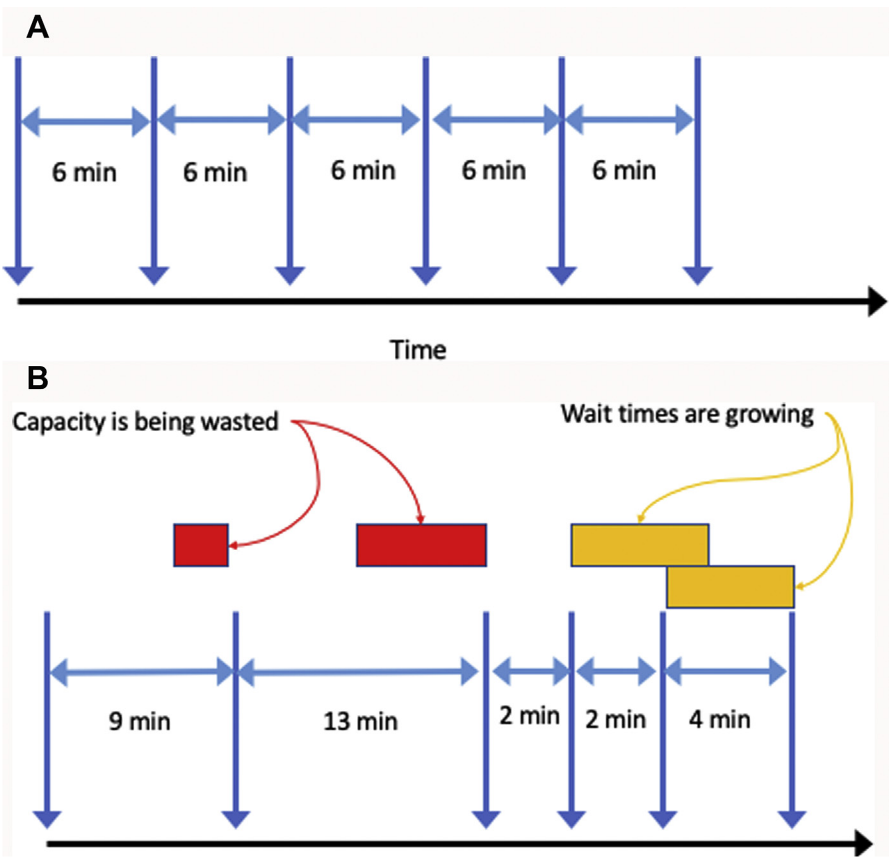

Fig. 1. (*A*) Regularly spaced arrivals; (*B*) Arrivals with typical variability.

they might be able to do something useful with that time, such as taking a break, going to the bathroom, or checking in with their colleagues. However, the only thing that they cannot do with the time they spend waiting to see a new patient is to use it to make up for future waits. There is no way to apply that time toward the patients who arrive when the nurses are busy; the amount of time they spend waiting simply accumulates, and can never be paid down.

When the variation between patient arrivals increases, the likelihood that patients will wait grows, as does the average amount of time they will wait. Although adding a third nurse at triage could certainly reduce these waits, doing so does not eliminate the potential for waits; it simply reduces them, because there could always be four patients who arrive in quick succession.

We can continue this process further as we add additional nurses, but doing so yields rapidly diminishing returns. Each additional nurse at triage is likely to spend a larger portion of their time waiting for a patient, because they will only be needed if a patient arrives while the preceding nurses are already busy. Using this strategy to staff our imaginary ED, which always sees six new patients an hour, means that the only way to ensure that patients never wait is to have six nurses at triage, one of whom will never see a patient unless all six of this hour's patients arrive in the same 10-minute period.

This becomes even less tenable once we start dealing with more realistic numbers of patients. If our six patients per hour really represents the average of a range from 1 patient per hour to 13, the potential costs of overstaffing become more glaring. Hospital management may want a "no wait" ED, but this is going to be a hard sell; we are better off trying to find if we can limit the amount of waiting to a degree that is feasible to staff and is still safe for our patients.

QUEUEING THEORY

What Is Queueing Theory?

Many of the pitfalls and tradeoffs that arise within these models are explained by queueing theory. Queueing theory is a branch of applied mathematics that is used to predict the behavior of lines (also known as queues), such as the length of a line over time, and the average time one spends waiting in a line. Although much of modern queueing theory emerged from the work of the Dutch mathematician A.K. Erlang, who used queueing models to predict the demand and use of telephone lines, it turns out that many of these properties apply no matter what the line is for. Queueing theory has been successfully applied across industries, streamlining services from manufacturing to airlines to restaurant service.[5–7]

Within the context of the ED, the same methodology is applied to model the queues for many different resources. Although there are slightly different assumptions that we need to make based on the kind of resource that we are modeling, the ways that the lines grow or shrink do not change. It does not matter whether the queues that we are concerned with are for physical resources, such as a CT scanner or laboratory tests, or for human resources, such as for registration clerks or for available nurses.

The mathematics involved in queueing analyses are sophisticated, but many of the lessons gained from even simple queueing models are useful to a broad variety of applications. You do not need to know the intricate details of queueing theory to reap many of its benefits. Furthermore, in day-to-day practice, many queueing models are approximated through the use of computer simulations and other programs that allow users to examine the behavior of underlying queues and the effects of changing their parameters, but without requiring the user to do the calculations themselves.

Little's Law

One of the most elementary theorems of queueing theory is Little's law.[8] Little's law states that the average size of a queue is proportional to the average rate of arrivals (usually referred to by the Greek letter λ), and the amount of time that a person spends in the queue.

L (average size of the queue) = λ (effective arrival rate) $\times$ W (average length of stay)

If an ED has an average arrival rate of six patients per hour, and the average patient has a length of stay of 3 hours, then Little's law dictates that the average occupancy of the ED will be 18 patients.[7,9]

Although Little's law is a fairly simple and intuitive equation, it actually took many years to prove conclusively. Little's law is a useful starting point for a queueing analysis because the formula is agnostic to how arrivals are distributed. If the average comes from patients being evenly distributed over time, or concentrated at any particular time interval (ie, if the patients mostly come at the end of the hour), then it still holds. We can also use Little's law for any portion of a larger system (we can apply Little's law to the ED directly, or we can apply it to any queue within the ED); it holds equally to describe the average occupancy of the ED, the average number of patients waiting for a CT scan, or any step of care that we choose.

Utilization

A direct result of Little's law, which provides a helpful understanding of the demand for a resource, is its utilization rate, commonly referred to by the Greek letter ρ. The utilization rate is the average patient arrival rate (λ) divided by the average service time (μ) for the resource. For instance, if a CT scanner receives four orders for CT scans per hour, and it can perform five scans per hour, then ρ of the CT scanner.

$\lambda/\mu = \rho$
is
$4/5 = 0.8$.

Although the utilization rate seems to only be a rearrangement of the terms from Little's law, it is useful because it provides a clear ratio of demand to capacity. If the utilization rate exceeds 1 at any point in time, then a backlog is guaranteed to develop and to continue getting worse. When accounting for increases in the variability of arrivals and service times, the ratio of utilization at which waits begin to grow rapidly may be considerably lower than one.

Kendall's Notation and Basic Queueing Models

Individual queues, which are used to examine any process within the ED, are frequently described using Kendall's notation.[5,10] Kendall's notation, typically in the form A/S/c, is a shorthand for describing the assumptions that underlie a queueing model. The "A" describes the arrival process (the pattern in which we assume patients arrive), "S" describes the service distribution (the pattern of how long it takes to provide a patient with a resource), and "c" refers to the number of servers (how many units of the resource we have). Fitting processes within the ED to these queueing models is helpful because some of them have closed form solutions, formulas that can reliably characterize how a process will perform over time.

A deterministic distribution, which has the notation "D," describes a process that occurs at a fixed rate. This is a reasonable description of some service distributions, but is generally a poor model for an arrival process. As we saw with the task of staffing triage, it is not particularly helpful to assume that patients arrive at a fixed rate.

However, some resources come close to always taking the same amount of time, such as a point-of-care blood glucose analyzer, which might be calibrated to take exactly 60 seconds for every sample. In Kendall's notation, our initial example, in which patients arrived at a fixed rate, and were triaged by two nurses who always took the same amount of time, would be known as a D/D/2 queue, because patients arrive at a determined rate, are served at a determined rate, and there are two nurses who provide the triage service.

A Markovian arrival process, which has the notation "M," has been shown in several studies to closely follow the arrival of the distribution of patient arrivals within EDs.[11–13] The most basic version of this process, the Poisson process, describes the time until the next patient arrives as an exponentially decreasing probability (**Fig. 2**). As a result, the average time between patient arrivals is often the balance point between "bursty" periods, where the interarrival times are short and several patients arrive in quick succession, and occasional long periods with no arrivals.

The simplest queue using this more rigorous model is the M/D/1 queue, which signifies patients arriving according to a Markovian arrival process, served by a single resource that follows a deterministic distribution. This model gives us a direct formula for the average time that a patient will need to wait in queue before they have access to the resource, which is:

$$\text{Waiting time} = \rho / 2\mu (1-\rho)$$

If we imagine a single blood gas analyzer, which receives a new sample on average every 6 minutes (average arrival time λ = 10 specimen arrivals per hour), and takes 2 minutes to perform its analysis (average service time μ = 30 specimens analyzed per hour), with a utilization of ρ = 1/3, then our formula suggests that a specimen will sit for about 5 minutes before it is analyzed.

The M/M/1 and M/M/c queue models expand on this to reflect variation in the amount of time it takes to deliver a service. This makes it a much more suitable model for triage, because we can imagine that the average amount of time to perform triage reflects some cases where the process is quick and straightforward (an ankle sprain), virtually instantaneous (a patient with a prehospital ST-segment elevation myocardial infarction alert who is triaged at the bedside), or potentially long (a patient requiring a translator who cannot recall why they came to the ED). Accounting for this variation in the service effectively doubles our estimated wait time:

$$\text{Waiting time} = \rho / \mu (1-\rho)$$

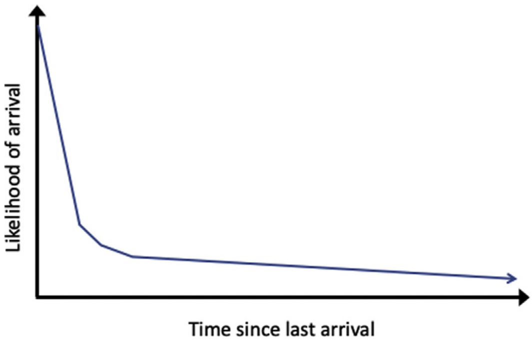

Fig. 2. The Poisson process graph.

Returning to our initial example of triage, with patients arriving every 6 minutes (λ = 10 patient arrivals per hour), and two nurses who each take on average 5 minutes to perform triage (μ = 12 patients triaged per hour), leading to a utilization of ρ = 10/12, then our formula predicts that patients will wait an average of 25 minutes before they begin being triaged. If an extra patient comes every hour, our utilization increases to ρ = 11/12, and our average wait time before being triaged jumps to a painful 66 minutes.

As we approach a utilization of 1 in the M/M/c model, the average wait grows quickly (Fig. 3), without an appreciable bound. That the point of inflection in the curve shifts as the number of a resource increases is a direct result of the degree to which utilization changes.[14] If there are three nurses at triage rather than two, then every additional patient who arrives represents a proportionally smaller increase in utilization. Thus, when allocating a resource based on its expected utilization, it is extremely important to pay attention not only to the average utilization of the resource, but to how large surges in demand can be.

The M/D/1 and M/M/c models can also be used to calculate the likelihood of a certain number of patients being within a queue at a point in time, and the probability that a patient in the queue will wait for more or less than a specific amount of time.

RELATING QUEUEING THEORY TO PRACTICE

Ways to Get Around Increased Utilization

Many EDs experience periods when utilization exceeds a ratio of 1, and waits begin to spiral out of control. At those times, waits are typically reduced in a few different ways. The most problematic one is that initiated by patients: leaving without being seen. Although there are many factors that affect patients' decision to leave without being seen, the rate at which they do so is modeled using a queueing formula based on service rates in the ED as a whole.[15] Although the previously mentioned queueing formulas assume no upper limit on the number of potential patients waiting, more complex models exist that can account for the effects of limited capacity, and patients' decreasing willingness to wait as delays mount.

Some EDs have successfully addressed burgeoning waits by allowing flexible staffing patterns and care areas. In many EDs, some of these staffing policies exist on an informal basis. If triage is swamped, and few patients are being brought into treatment areas, more nurses may go to triage to help relieve their colleagues. In other EDs, dedicated fast track areas can be opened at periods of high demand, particularly if there are times of day in which there are higher volumes of patients with low-acuity presentations.[4,15–17]

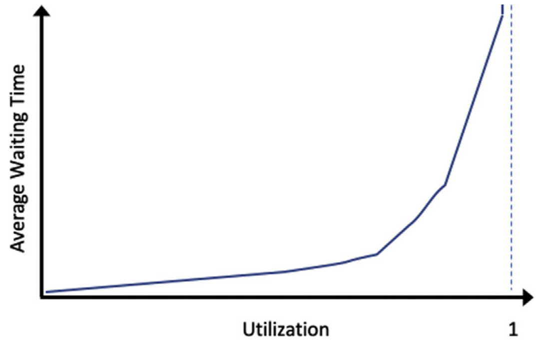

Fig. 3. Average wait time in the M/M/c model.

Similarly, some EDs have experimented with having lower-acuity patients schedule appointments within fast-track areas. Although the concept of an ED "appointment" may seem like a contradiction in terms, for urgent, but not necessarily emergent conditions that are not easily handled in a primary care office (eg, lacerations), this practice has merit. From a queueing theory standpoint, having patients schedule appointments, even if it is only an hour or two in advance, effectively turns their arrivals into a deterministic process, which can substantially reduce the amount of time they ultimately wait in queue.

A Successful Application of Queueing Theory in the Emergency Department

Perhaps the most successful application of queueing theory to ED operations was undertaken by Green and colleagues,[16] who used a series of M/M/c models to match the number of EPs on staff in a community hospital ED to match arrival rates at different times of day, factoring in a delay between the peak arrival rate in a given time period and the point at which it reached peak congestion. The study found that on days on which there was no change in the total number of physician hours, and hours were simply redistributed to better align with demand, there were substantial decreases in the number of patients who left without being seen, despite an increase in the overall volume of patients seen.

Can Getting Rid of the Waiting Room Speed up the Emergency Department?

The queueing models provide a mechanistic explanation for why waiting cannot be meaningfully eliminated in the ED, and unless an ED plans to let patients waiting to be seen line up outside, some kind of waiting room must remain. However, to the extent that "eliminating the waiting room," often refers to posting a physician at triage, it is an effective means of decreasing rates of patients leaving without being seen, and can decrease patients' overall length of stay.[17–19] Furthermore, the time patients spend in the waiting room before seeing a physician is often the kind of delay that they find most intolerable, making reducing this time a priority, even if overall delays are difficult to reduce.[20–22]

Posting a physician at triage is most appropriate for EDs that identify significant delays between patients completing triage and are brought back to a treatment space and evaluated by a physician. This is because of delays in patient transport, but is most commonly caused by crowding; a patient cannot be brought back if there is nowhere to bring them. This scenario is identified quickly by comparing the utilization of EPs relative to the utilization of triage. If the utilization of EPs is low, then moving the physician evaluation earlier in the process is probably worthwhile. If the utilization of EPs is already high, then moving physicians into triage may still reduce the left without being seen rate, but may not affect or even worsen some patients' length of stay.

Applying Queueing Theory to Everyday Emergency Department Operations

The performance of the ED itself reflects the interaction of multiple interlocking queues. A patient's evaluation and treatment will progress through triage and physician assessment, and can end there, or it can depend on several more queues, such as for laboratory tests, a CT scan, and an inpatient bed. The equations for the M/D/1 and M/M/c queues are powerful tools that can be used to analyze these processes and locate potential bottlenecks, which can cause a decrease in the utilization of processes elsewhere in the system. Multiple free online tools are available for analyzing basic queue models, including average waiting times and queues' probable occupancy (eg, https://www.supositorio.com/rcalc/rcalclite.htm and https://www.edqueue.org).

However, for the ED as a whole, no single set of equations can reliably predict the effects of changing a process on the system itself. Instead, software simulations are used to factor in the behavior of multiple queues over many trials and examine the effects of individual changes over time. Often, the process of improving the movement of a single queue within the ED, such as laboratory testing or imaging, can have significant downstream effects that simulation can effectively predict.[23–26] Simulation frameworks for queueing models are available as bespoke commercial applications (eg, ARENA, ProSim),[27] through packages for commercial software (eg, MATLAB), and as libraries for open-source software (eg, for C++ and Python).[28] Packages also exist for performing queueing analyses and simulations using commonly available spreadsheet software (eg, Microsoft Excel).

The prospect of modeling the operations of an entire ED may seem daunting, but the potential payoffs from applying queueing theory are immense. Occasionally a single bottleneck is critical to increased waits throughout the ED, even for patients who do not require its resources.[23] Although there exists no single solution to streamlining an ED, the application of queueing theory holds keys to identifying the most valuable targets for improvement, which remain unexplored in practice, despite a growing literature of simulation frameworks.[29–31] Whether these gains will ever be realized, however, remains in the hands of individual administrators.

DISCLOSURE

The author has nothing to disclose.

REFERENCES

1. Box GEP, Draper NR. Empirical model-building and response surfaces. Oxford (England): John Wiley & Sons; 1987. xiv, 669. (Empirical model-building and response surfaces).Introduction.
2. FitzGerald G, Jelinek GA, Scott D, et al. Emergency department triage revisited. Emerg Med J 2010;27(2):86–92.
3. Twomey M, Wallis LA, Myers JE. Limitations in validating emergency department triage scales. Emerg Med J 2007;24(7):477–9.
4. Wiler JL, Gentle C, Halfpenny JM, et al. Optimizing emergency department front-end operations. Ann Emerg Med 2010;55(2):142–60.e1.
5. Shortle JF, Thompson JM, Gross D, et al. Fundamentals of queueing theory. Hoboken, NJ: John Wiley & Sons; 2018. p. 566.
6. Takagi H. From computer science to service science: queues with human customers and servers. Comput Netw 2014;66:102–11.
7. Harris M. Little's law: the science behind proper staffing [Internet]. Emergency Physicians Monthly. 2010. Available at: https://epmonthly.com/article/littles-law-the-science-behind-proper-staffing/. Accessed November 17, 2019.
8. Little JDC. Little's law as viewed on its 50th anniversary. Oper Res 2011;59(3): 536–49.
9. Lovejoy WS, Desmond JS. Little's law flow analysis of observation unit impact and sizing. Acad Emerg Med 2011;18(2):183–9.
10. Kendall DG. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann Math Stat 1953; 24(3):338–54.
11. Armony M, Israelit S, Mandelbaum A, et al. On patient flow in hospitals: a data-based queueing-science perspective. Stoch Syst 2015;5(1):146–94.

12. Whitt W, Zhang X. A data-driven model of an emergency department. Oper Res Health Care 2017;12:1–15.
13. Kim S-H, Whitt W. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manuf Serv Oper Manag 2014;16(3):464–80.
14. Millsap C. Thinking clearly about performance. Queue 2010;8(9):10.
15. Wiler JL, Bolandifar E, Griffey RT, et al. An emergency department patient flow model based on queueing theory principles. Acad Emerg Med 2013;20(9): 939–46.
16. Green LV, Soares J, Giglio JF, et al. Using queueing theory to increase the effectiveness of emergency department provider staffing. Acad Emerg Med 2006; 13(1):61–8.
17. Abdulwahid MA, Booth A, Kuczawski M, et al. The impact of senior doctor assessment at triage on emergency department performance measures: systematic review and meta-analysis of comparative studies. Emerg Med J 2016;33(7):504–13.
18. Imperato J, Morris DS, Binder D, et al. Physician in triage improves emergency department patient throughput. Intern Emerg Med 2012;7(5):457–62.
19. Zayas-Caban G, Xie J, Green LV, et al. Policies for physician allocation to triage and treatment in emergency departments. IISE Trans Healthc Syst Eng 2019;0(0):1–15.
20. Maister D. The psychology of waiting lines. In: Czepiel J, Solomon M, Surprenant C, editors. The service encounter. Lexington (MA): Lexington Books; 1985. p. 113–23.
21. Soremekun OA, Takayesu JK, Bohan SJ. Framework for analyzing wait times and other factors that impact patient satisfaction in the emergency department. J Emerg Med 2011;41(6):686–92.
22. Dahlen I, Westin L, Adolfsson A. Experience of being a low priority patient during waiting time at an emergency department. Psychol Res Behav Manag 2012;5:1–9.
23. Storrow AB, Zhou C, Gaddis G, et al. Decreasing lab turnaround time improves emergency department throughput and decreases emergency medical services diversion: a simulation model. Acad Emerg Med 2008;15(11):1130–5.
24. Zeng Z, Ma X, Hu Y, et al. A simulation study to improve quality of care in the emergency department of a community hospital. J Emerg Nurs 2012;38(4):322–8.
25. Paul JA, Lin L. Models for improving patient throughput and waiting at hospital emergency departments. J Emerg Med 2012;43(6):1119–26.
26. Hung GR, Whitehouse SR, O'Neill C, et al. Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. Pediatr Emerg Care 2007;23(1):5–10.
27. Wang T, Guinet A, Belaidi A, et al. Modelling and simulation of emergency services with ARIS and Arena. Case study: the emergency department of Saint Joseph and Saint Luc Hospital. Prod Plan Control 2009;20(6):484–95.
28. Dagkakis G, Heavey C. A review of open source discrete event simulation software for operations research. J Simul 2016;10(3):193–206.
29. Salmon A, Rachuba S, Briscoe S, et al. A structured literature review of simulation modelling applied to emergency departments: current patterns and emerging trends. Oper Res Health Care 2018;19:1–13.
30. Mohiuddin S, Busby J, Savović J, et al. Patient flow within UK emergency departments: a systematic review of the use of computer simulation modelling methods. BMJ Open 2017;7(5):e015007. Available at: https://bmjopen.bmj.com/content/7/5/e015007. Accessed November 15, 2019.
31. 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.