Five easy equations for patient flow through an emergency department

Thomas Lill Madsen[1] & Allan Kofoed-Enevoldsen[2]

ABSTRACT
INTRODUCTION: Queue models are effective tools for framing management decisions and Danish hospitals could benefit from awareness of such models. Currently, as emergency departments (ED) are under reorganization, we deem it timely to empirically investigate the applicability of the standard "M/M/1" queue model in order to document its relevance.
MATERIAL AND METHODS: We compared actual versus theoretical distributions of hourly patient flow from 27,000 patient cases seen at Frederiksberg Hospital's ED. Formulating equations for arrivals and capacity, we wrote and tested a five equation simulation model.
RESULTS: The Poisson distribution fitted arrivals with an hour-of-the-day specific parameter. Treatment times exceeding 15 minutes were well-described by an exponential distribution. The ED can be modelled as a black box with an hourly capacity that can be estimated either as admissions per hour when the ED operates full hilt Poisson distribution or from the linear dependency of waiting times on queue number. The results show that our ED capacity is surprisingly constant despite variations in staffing. These findings led to the formulation of a model giving a compact framework for assessing the behaviour of the ED under different assumptions about opening hours, capacity and workload.
CONCLUSION: The M/M/1 almost perfectly fits our ED. Thus modeling and simulations have contributed to the management process.
FUNDING: not relevant.
TRIAL REGISTRATION: not relevant.

The literature describing queue models for the emergency department (ED) is substantial. The models encourage the practitioner to balance demand for service and supply of facilities in a systematic and cost-effective manner [1]. However, to the best of our knowledge, the models are underutilized in the Danish context and our assumption is that this may apply across the board to other countries. The relative complexity of the models may act as a barrier for their utilization.

In this paper, we addressed the potential for better management. For a specific ED, we showed how well a simple queue model worked and how relatively easy it may be to achieve better management through the use of such a model.

Queue models generally deal with customer arrivals at a service facility. The main parameters are the arrival pattern of customers, how much service they need and the capacity of the server. In its most basic form, this is referred to as the "M/M/1" model. The notation implies here that 1) customer arrivals are generated by a Poisson process, 2) their service demand follows an exponential distribution and 3) a single server does the processing. 1) and 2) are also known as Markov processes, i.e. a system characterized completely by its current state, regardless how it arrived at such state. This type of model assumptions has been widely used as representations of telephone centrals, computer systems and other traffic carrying entities with independent arrivals.

We 1) analyzed the patient flow for an entire year at Frederiksberg Hospital's ED and compared the empirical and theoretical distributions of arrival and treatment times. And 2), on this basis, we proceeded to write a small system of equations that simulate the ED, 3) checked consistency with observations and 4) showed how this model may be used to schedule staffing.

MATERIAL AND METHODS

From March 2009 to February 2010, a total of 27,142 arrivals were recorded. For each arrival we know the three event times: time of arrival (t0), time of admission (t1) and time of discharge (t2). Waiting time is t1-t0 and service time is t2-t1. From these data we were able to reconstruct the queue at a given time and to calculate the queue number for each patient. By definition, service times were zero for patients who reneged, 7%, and others, 3%. Patients who reneged are relatively more frequent in the latter half of the afternoon and evening when the queue is long.

For about 4,000 patients, admission time (t1) was missing. Rather than excluding these cases, which would lead to a number of problems, t1 was generated artificially as t2-u, where u is exponentially distributed with mean = 60 minutes under the constraint that t1 > t0. The justification for this procedure will become apparent in due course.

Trial registration: not relevant.

ORIGINAL ARTICLE

1) The Danish National Centre for Social Research and
2) Department of Medicine, Frederiksberg Hospital

Dan Med Bul 2011;58(10):A4318

Emergency department waiting room.

RESULTS

The ED is driven by the processes of arrival, admission and discharge. On average the queue resets to zero in the early hours of the morning and rises to eight at noon remaining stable until 8 p.m. Thereafter, until the following morning, admissions prevail over arrivals and the queue goes to zero. On arrival, the average patient is confronted with a queue of 6.6 (**Figure 1**).

Arrivals (M/M/1)

The M/M/1 model assumption is that arrivals per time unit numerically behave as drawings from a Poisson distribution. The number of arrivals varies by a minimum of three time dimensions: month, weekday and hour. We found that hourly variation is the absolute dominant component and as a first approximation we could ignore variation from both weekday and month. We found no significant relation between arrivals from one hour to the next, confirming the Markov property.

For the number of arrivals by the hour, we compared the empirical distribution to what was expected from Poisson distributions with the same means. This showed a very good correspondence, even without taking seasonality into account. We have thus found that the number of arrivals for each hour of the day have a *standard deviation* of almost exactly the same magnitude as the square root of the *mean,* which is the signature of a Poisson process.

The clinical implication of this is clear: if on average you expect n patients to arrive during any hour of the day, from the Poisson distribution with the mean n you can immediately look up the probability of receiving any specific number of patients. Another corollary is that this rule favours larger medical units over small: the relative magnitude of randomness diminishes with size and allows for better utilization of inputs.

The arrival component of our ED-model is then simplified:

A(t) = Ps(a(t))

Where A(t) is the simulated number of arrivals in hour "t", a(t) is the average intensity in hour "t" and Ps(x) is a drawing from the Poisson distribution with mean x.

Service times (M/M/1)

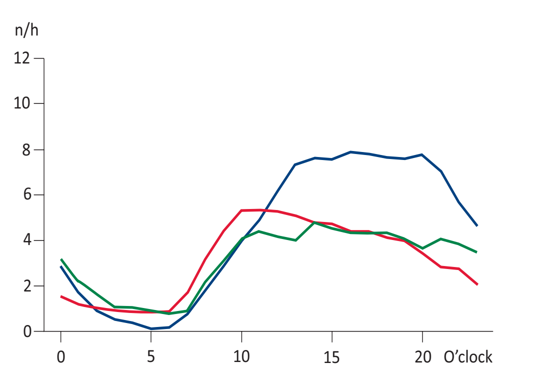
The average treatment time is 61 minutes, excluding reneges (who do not come into contact with medical staff and thus contribute no information about service times and patients with treatment times exceeding five hours (n = 721). The standard hypothesis is that service times obey an exponential distribution. For times not exceeding 15 minutes this does not hold as there are too few of these. However, for treatment times exceeding 15 minutes, we found that frequencies exhibit a perfect exponential decay with a parameter of 0.0185, meaning an average extra service time of 1/0.0185 = 54 minutes (in excess of the initial 15 minutes). The clinical insight this contributes is the following: for a randomly chosen patient at a randomly chosen time, there is a 1.85% chance that the treatment will be completed during the next minute, no matter how long it has already lasted. Also, service times apparently form a perfect continuum with no natural subdivisions.

The server (M/M/1)

M/M/1 postulates a single server, e.g. a single operator in call center. Our ED is staffed with either one or two servicing teams each with one physician (MD) and two nurse practitioners (NP); thus, a complex entity. Nevertheless, for modelling purposes, it turns out that we can think of this unit as a black box with a certain capacity

FIGURE 1

Average arrivals, admissions and resulting queue length per hour over the day.



measured by how many average patients can be seen during one standard hour. Ideally, capacity should be proportional to the medical staff assigned and inversely proportional to treatment times. Further, since service times differ randomly as described, capacity must necessarily be a stochastic quantity with a certain mean and standard deviation. Further, there may conceivably be times of the day where the staff productivity deviates from average.

Without going into details, we would expect the Poisson distribution to be a good model for this black box and have verified this by comparing the distribution of "number of admitted patients per hour" over a busy time interval of the day to the Poisson distribution with the same mean = 4.1. We have thus found the Poisson distribution to be a good approximation of the server's behaviour at full utilization (**Figure 2**).

We thus have one further model component:

C(t) = Ps(c(t))

where C(t) is the simulated number of possible admissions in hour "t" and c(t) is the average achievable capacity in hour "t".

A regression approach for capacity estimation can be based on individual waiting times (w) for different queue lengths. The relation for this is w = Q/C where the patient waiting next in line is assigned queue no. 1 (data for Q = 1 have to be omitted since this is a special case: short waiting times arise due to situations where the patient arrives when the ED is partially idle and the newly arrived patient is treated instantly. Capacity can therefore be calculated by regressing waiting time on queue length. Now, there are two different meanings to "queue length" for a given patient: a) the total number of patients in the waiting room at the time of arrival of the index patient and b) the number of patients that actually were seen from the time of arrival to his time of admission. We will be using the latter definition as it reflects more accurately on waiting time.

Variations in staffing level and other factors may cause capacity to change during the day. For each case (n = 23,000, excluding the 2.5% highest values of Q), we therefore regress waiting times on queue lengths in interaction with the hour of the day. Solved as a linear model with a square root transformation of w and Q (to avoid heteroscedaticity), 24 coefficients are estimated (the coefficients attained include renege implicitly and are therefore about 10% higher than the 4.1 net value of mentioned earlier). This is shown in **Figure 3**.

The regression yields a rather constant capacity over the day with an average value of 4.6 patients/hour. I.e. if you are number four in line for service, you will be waiting about 50-60 minutes regardless of the hour. A quite surprising finding since the staffing during the night is only one team versus two teams otherwise.

The simulation model

Putting it all together, the components for a 5-equation simulation model of the ED are now in place:

for t = 0 to 23

(1)	A(t) = Ps(a(t))	Arrivals in hour "t"
(2)	Q(t) = Q(t-1) + A(t)	Preliminary queue
(3)	C(t) = Ps(c(t))	Capacity potential
(4)	X(t) = min(Q(t),C(t))	Completed treatments
(5)	Q(t) = Q(t)-X(t)	Resulting queue ultimo "t"

next t

Repeat loop 365 times to simulate a full year.

The performance of the model (including an estimated hourly capacity of 4.6 for all time periods except t = 7 where the capacity was estimated to three as de-

FIGURE 2

Distribution of treatment times. Treatment times longer than 15 minutes follow an exponential distribution.

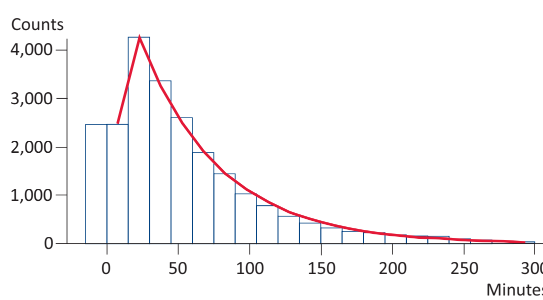


FIGURE 3

Regression estimate of hourly capacity.

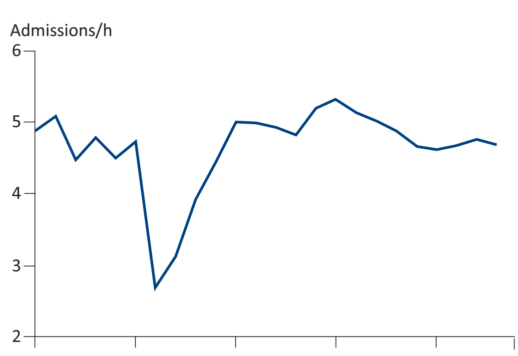
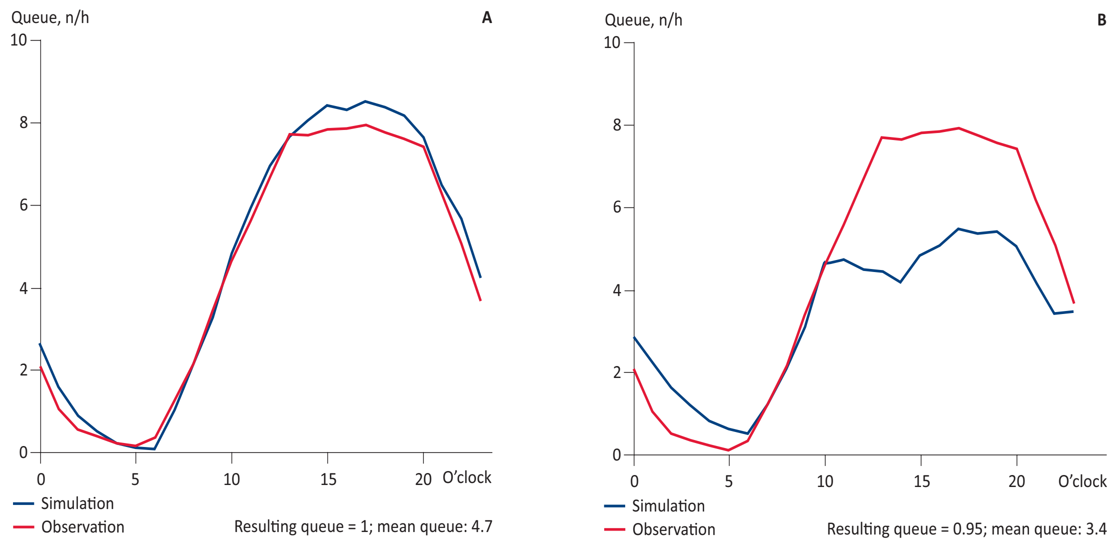


FIGURE 4



Model test and simulation. **A.** Model performance, 365 days aggregate. **B.** Simulation: What if we move two units of capacity from the night shift to the time interval 11:00-15:00?

rived below) is demonstrated in **Figure 4** taking queue length as a representative variable for the system.

Modelling patient flow, we are able to extract critical performance measures such as capacity per hour and estimated waiting times as well as simulating queue length in scenarios involving modified capacity. We find the distributional properties of the simulated queue length to be quite similar to the observed queue length.

We finally simulated the effect of moving two capacity units from the night shift to the later part of the day shift. As demonstrated in Figure 4B, the average queue then dropped from 4.7 to 3.4 patients, implying a 27% reduction in waiting time and, most significantly, in the afternoon where waiting times are too long.

This was mostly a theoretical exercise since the night staff (1 MP + 2 NP) is thought not to be reducible; nevertheless, it illustrates the type of problem the model can deal with.

DISCUSSION

Using a large dataset, we found the M/M/1 to be a good approximation of patient flow in this ED, including the ability of the server to admit and discharge patients analogous to carried traffic [2]. Arrivals, service times and server capacity were all found to lie within the realm of the exponential family of distributions. The M/M/1 model may therefore constitute the basis for a compact way of simulating the ED and we have here demonstrated its potential.

Interestingly, our model is equivalent to a discrete event model of the same system, but definitely easier to explain and program. Other models in which the ED is analyzed as a number of stages or sub processes may offer more detailed insight and sophisticated simulation opportunities at the expense of input information that cannot be as easily obtained [3, 4].

It is our guess that many ED suppliers are unaware of the actual size distribution of their hourly capacity. In our model case, a surprising uniformity of capacity over the day comes to light which raises fundamental questions about the relation between staffing schedule and throughput which again identifies possible inefficiencies of the day shift. Our ED is manned by two teams of each one MD and two NP during the day (8 a.m.-18 p.m.) and one such team during the evening and night shift. The information about capacity uniformity has changed our focus from a wish for more hands towards work flow redesign.

The model and our main findings were communicated to the ED staff and discussed at staff meetings. The initial counter-intuitive finding of a lack of correlation between staffing and server capacity was well-received and has led to a prompt shift of focus from a claim for "more hands" to the design of more effective organization of work flow. A number of noise-creating factors such as student supervision as well as physical and communicative environmental barriers have been identified and innovative concepts such as creating a parallel fast track line and rebuilding of the reception area have been initiated. In addition, the possibility of communicating the expected waiting time to incoming patients is currently being explored [5]. Besides creating a common understanding of the foundations of patient flow dynamics, the possibility of visualizing the effects of a potential capacity increase on queue length has enabled a clear and easily communicable vision of "halving the afternoon queue by increasing our daytime capacity with two patients per hour". The monitoring of our forthcoming performance is expected to be markedly facilitated by analyzing future patient flow data using our model.

There is no reason to assume that other EDs are fundamentally different to the one examined herein. It is therefore our hope that this study will draw suppliers' attention to the demonstrated opportunities of improved system awareness, management and service.

CORRESPONDENCE: *Thomas Lill Madsen,* Sagasvej 18, 1861 Frederiksberg C, Denmark. E-mail: thomas.lilll.madsen@gmail.com

ACCEPTED: 12 July 2011

CONFLICTS OF INTEREST: none

ACKNOWLEDGEMENT: Statistical and graphical software: R Development Core Team (2006). R: A language and environment for statistical computing. Vienna (Austria), R Foundation for Statistical Computing.

LITERATURE

1. Channouf N, L'Ecuyer P, Ingolfsson A et al. The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. Health Care Manag Sci 2007;10:1025-45.
2. http://en.wikipedia.org/wiki/Erlang_(unit) (1 Oct 2010).
3. Mayhew L, Smith D. Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments. Health Care Manag Sci 2008;11:11-21.
4. Au-Yeung SWM, Harrison PG, Knottenbelt WJ. A queuing network model of patient flow in an accident and emergency department, 2006 http://pubs.doc.ic.ac.uk/aande-queueing-model/aande-queueing-model.pdf (3 Aug 2011).
5. Madsen TL. Klar besked om ventetid. Sygeplejersken 2010;15:36.