Queuing Analytic Theory and Discrete Events Simulation for Healthcare: Right Application for the Right Problem

Alexander Kolker, PhD
Outcomes Operation Project manager
Children's Hospital of Wisconsin

Introduction

Modern healthcare achieved great progress in developing and manufacturing medical devices, treatment procedures and drugs to serve individual patients. However, relatively little technical talent and material resources have been devoted to improving the overall health care delivery process as a system.

Discrete-event simulation (DES) models and queuing analytic (QA) theory are the most widely applied system engineering and operations research methods used for system analysis and justification of operational business decisions.

There are some proponents of using QA theory to solve many pressing hospital problems of patient flow and variability, calculating needed nursing resources and the number of beds and operating rooms (IHI, 2008; Litvak, 2007; McManus et al, 2004; Haraden et al, 2003). They 'prescribe' QA theory as a necessary part of a so-called managing variability program (MVP) (IHI, 2008). However they tend to ignore some serious practical limitations of QA theory for hospitals applications. D'Alesandro (2008) summarized why QA theory is misplaced in hospitals.

On the other hand, there is a strong tendency of using discrete-events simulation (DES) as a preferred and much more versatile methodology of health care management science to address complex problems of patient flow and predicting needed resources. Jacobson et al (2006) presented a review of publications since 1999 on DES of health care systems.

The objective of this material is to provide a comparative analysis of both QA and DES models by applying them side by side to the same problems. Using a number of concrete quantitative examples it is demonstrated why QA approach has serious practical limitations while DES is indeed a more powerful, flexible and informative methodology than QA.

Queuing analytic models: their use and limitations

The term 'queuing theory' is usually used to define a set of analytic techniques in the form of closed mathematical formulas to describe properties of the processes with a random demand and supply (waiting lines or queues). Queuing formulas are usually applied to a limited number of pre-determined simplified models of the real processes for which analytic formulas can be developed.

Weber (2006) cites that '…There are probably 40 (queuing) models based on different queue management goals and service conditions…' and that it is easy '… to apply the wrong model' if one does not have a strong background in operations research.

Development of tractable analytic formulas is possible only if a flow of events in the system is a steady-state Poisson processes. The latter is, on definition, an ordinary stochastic process of independent events with the parameter equal to the intensity of the corresponding flow. Time intervals between events in a Poisson flow are exponentially distributed with the average inter-arrival time that is inverse to a Poisson intensity parameter. Service time is assumed to follow an exponential distribution or, sometimes, uniform, Erlang or some distributions with the coefficient of variation close to 1 (Green, 2006). Thus, processes with a Poisson arrival of events (exponential inter-arrival time) and exponential service time are Markov stochastic processes with discrete states and continuous time.

Most widely used queuing models for which relatively simple closed analytical formulas have been developed are specified as *M/M/s* type (Hall, 1990; Lawrence and Pasternak, 1998). (*M* stands for Markov since Poisson process is a particular case of a stochastic process with no ‘after-effect’ or no memory, known as continuous time Markov process). These models assume an unlimited queue size that is served by *s* providers.

Typically *M/M/s* queuing models allow calculating the following steady-state characteristics (Lawrence and Pasternak, 1998):

- probability that there are zero customers in the system
- probability that there are *K* customers in the system
- the average number of customers waiting in the queue
- the average time the customers wait in the queue
- the average total time the customer spends in the system (‘cycle time’)
- utilization rate of servers , i.e. percentage of time the server is busy

As more complexity is added in the system, the analytic formulas become less and less tractable. Analytic formulas are available that include, for example, limited queue size, entities leaving the system after waiting a specified amount of time, multiple queues with different average service time and different providers’ types, different service priorities, etc (Lawrence and Pasternak, 1998; Hall, 1990). However the use of these cumbersome formulas even built in spreadsheets or tables (Hillier, Yu, 1981; Seelen et al, 1985) is not convenient and rather limited for most practical applications in healthcare.

Assumptions that allow deriving most queuing formulas are not always valid for many healthcare processes. For example, several patients sometimes arrive in Emergency Department at the same time (several people injured in the same auto accident), and/or the probability of new patient arrivals could depend on the previous arrivals when ED is close to its capacity, or the average arrival rate varies during a day, etc. These possibilities alone make the arrival process a non-ordinary, non-stationary with after-effect, i.e. a non-Poisson process for which queuing formulas are not valid. Therefore it is important to properly apply statistical goodness of fit tests to verify that the null-hypothesis that actual arrival data follow a Poisson distribution cannot be rejected at some level of significance.

An instructive example of not convincing conclusion from the goodness-of-fit statistical test is in Harrison et al (2005). These authors stated that ‘…it is valid to model all admissions on any particular day of the week as a Poisson process but the mean admission rates vary dramatically from day to day’. A Poisson process, however, could not approximate the entire week arrivals. Harrison et al (2005) tried to justify their conclusion by using a chi-square goodness of fit test. The authors obtained the test p-values in the range from 0.136 to 0.802 for different days of the week. Because p-values were greater than 0.05 level of significance, they failed to reject the null-hypothesis of Poisson distribution (accepted the null-hypothesis).

On the other hand, the fundamental property of a Poisson distribution is that its mean value is equal to its variance (squared standard deviation). It follows from the authors' own data that the mean value was not even approximately equal to the variance for Tuesdays (mean 27.15, variance 30.96), Thursdays (mean 27.73, variance 19.33), Fridays (mean 23.06, variance 32.01) and Saturdays (mean 20.64, variance 15.45). Thus, the use of a Poisson distribution was not actually convincingly justified for the patient arrivals on at least four days of the week. Apparently, chi-square test p-values were not large enough to accept the null-hypothesis with high enough confidence (alternatively, the power of the statistical test was likely too low) (Glantz, 2005)

Despite its rather limited applicability to many patient arrival patterns, a Poisson process is widely used in operation research as a standard assumption because of its mathematical convenience (Gallivan, 2002; Green, 2006; Green et al, 1991; McManus et al, 2003; McManus et al, 2004).

Some authors are trying to make queuing formulas applicable to real process by fitting and calibration. For example, in order to use queuing formulas for a rather complex ED system Mayhew and Smith (2008) made a significant process simplification by presenting the workflow as a series of stages. The stages could include initial triage, diagnostic tests, treatment, and discharge. Some patients experienced only one stage while others more than one. However, '… what constitutes a 'stage' is not always clear and can vary…and where one begins and ends may be blurred' (Mayhew and Smith, 2008). The authors assumed a Poisson arrival and exponential service time but then used actual distribution service time for 'calibration' purposes. Moreover, the authors observed that exponential service time for the various stages '…could not be adequately represented by the assumption that the service time distribution parameter was the same for each stage'. In the end, all the required calibrations, adjustments, fitting to the actual data made the model to lose its main advantage as a queuing model: its analytical simplicity and transparency. On the other hand, all queuing formulas assumptions and approximations still remained.

Therefore many complex healthcare systems with interactions and interdependencies of the subsystems cannot be effectively analyzed using analytically derived closed formulas.

Moreover, queuing formulas cannot be directly applied if the arrival flow contains a non-random component, such as scheduled arrivals (see examples below). Therefore, in order to use analytic queuing formulas, the non-random arrival component should be first eliminated leaving only random arrival flow for which QA formulas could be used (Litvak, 2007; Litvak and Long, 2000).

Green (2004) applied M/M/s model to predict delays in the cardiac and thoracic surgery unit with mostly elective scheduled surgical patients assuming a Poisson pattern of their arrivals. The author acknowledged that this assumption could result in an overestimate of delays. In order to justify the use of M/M/s model the author argued that some '…other factors are likely to more than compensate for this'. However, it was not clear what are those factors and how much and why they could compensate the overestimated delays (compare with example 5 in section 2.2.1).

Still, despite their limitations, QA models have some place in operation research for application to simply structured steady-state processes if a Poisson arrival and exponential service time assumptions are accurate enough.

DES models: basic applications

In contrast to queuing formulas, DES models are much more flexible and versatile. They are free from assumptions of the particular type of the arrival process (Poisson or not), as well as the service time (exponential or not). They can be used for the combined random and non-random arrival flow. The system structure (flow map) can be complex enough to reflect a real system structure, and custom action logic can be built in to capture the real system behavior.

At the same time it should be noted that building a complex realistic simulation model requires sometimes a significant amount of time for custom logic development, debugging, model validation, and input data collection.

However a good model is well worth the efforts because it becomes a powerful and practically the only real tool for a quantitative analysis of complex hospital operations and decision-making. Jacobson at al (2006) provided the latest review of DES applications for Healthcare delivery.

Moreover, many currently available simulation software packages (ProcessModel, ProModel, Arena, Simula8, and many others) provide a user-friendly interface that makes the efforts of building a realistic simulation model not more demanding than the efforts needed to make simplifications, adjustments and calibrations to develop a rather complex but inaccurate queuing model. Swain (2007), Abu-Taeh et al (2007), Hlupic (2000), Nikoukaran (1999) provided a review and a comparative study of dozens commercially available simulation packages.

A DES model is a computer model that mimics the dynamic behavior of a real process as it evolves with time in order to visualize and quantitatively analyze its performance. The validated and verified model is then used to study behavior of the original process and then identify the ways for its improvement (scenarios) based on some improvement criteria. This strategy is significantly different from the hypothesis-based clinical testing widely used in medical research (Kopach-Konrad et al, 2007).

Typical DES applications include: staff and production scheduling, capacity planning, productivity improvement, cycle time and cost reduction, throughput capability, resources and activities utilization, bottleneck finding and analysis. DES model is the most effective tool to perform quantitative 'what-if' analysis, and play different scenarios of the process behavior as its conditions and variables change with time. This simulation capability allows one to make experiments on the computer display, and to test different solutions (scenarios) for their effectiveness before going to the hospital floor for the actual implementation.

The basic elements (building blocks) of a simulation model are:

- Flow chart of the process, i.e. a diagram that depicts logical flow of a process from its inception to its completion
- Entities, i.e. items to be processed, e.g. patients, documents, customers, etc.
- Activities, i.e. tasks performed on entities, e.g. medical procedures, exams, document approval, customer check in, etc

- Resources, i.e. agents used to perform activities and move entities, e.g. service personnel, operators, equipment, nurses, physicians
- Entity routings that define directions and logical conditions flow for entities

Typical information usually required to populate the model includes:

- Quantity of entities and their arrival time, e.g. periodic, random, scheduled, daily pattern, etc. There is no restriction on the arrival distribution, such as Poisson distribution, required by the closed analytical formulas of the queuing theory
- The time that the entities spend in the activities, i.e. service time. This is usually not a fixed time but a statistical distribution. There is no restriction for a special exponential service time distribution required by the closed analytical formulas of the queuing theory
- The capacity of each activity, i.e. the max number of entities that can be processed concurrently in the activity.
- The maximum size of input and output queues for the activities
- Resource assignments: their quantity and availability, and / or working shift schedule

DES models work by tracking entities moving through the system at distinct points of time (events). DES records the detailed track of all processing times, waiting times, and gather entities' statistics at any activity and any point of time. DES is capable of tracking hundreds of individual events over the period of simulation time, each with its own unique attribute, enabling one to replicate the most complex systems with interacting components and interdependencies.

Analysis of patient flow is an example of the general dynamic supply and demand problem. There are three basic components that should be accounted for in such problems: (i) the number of patients (or, generally, any items) entering the system at any point of time, (ii) the number of patients (or any items) leaving the system after spending some time in it, (iii) capacity of the system which limits the flow of items through the system. All three components affect the flow of patients (items) that the system can handle. A lack of the proper balance between these components results in the system's over-flow and gridlock. DES methodology provides invaluable means for analyzing and managing the proper balance.

It will be demonstrated in the following sections that even simple DES models have a significant advantage over QA models. To illustrate this advantage DES methodology will be applied to the same processes that have been analyzed using QA.

Example 1

Flu clinic: unlimited queue size with steady state operation

A small busy clinic provides flu shots during a flu season on a walk-in basis (no appointment necessary). The clinic has two nurses (servers). Average patient arrival rate is 54 patients per hour with about the same number of elderly and all others. Each shot takes on average about 2 min.

It is required to calculate characteristics of the queue.

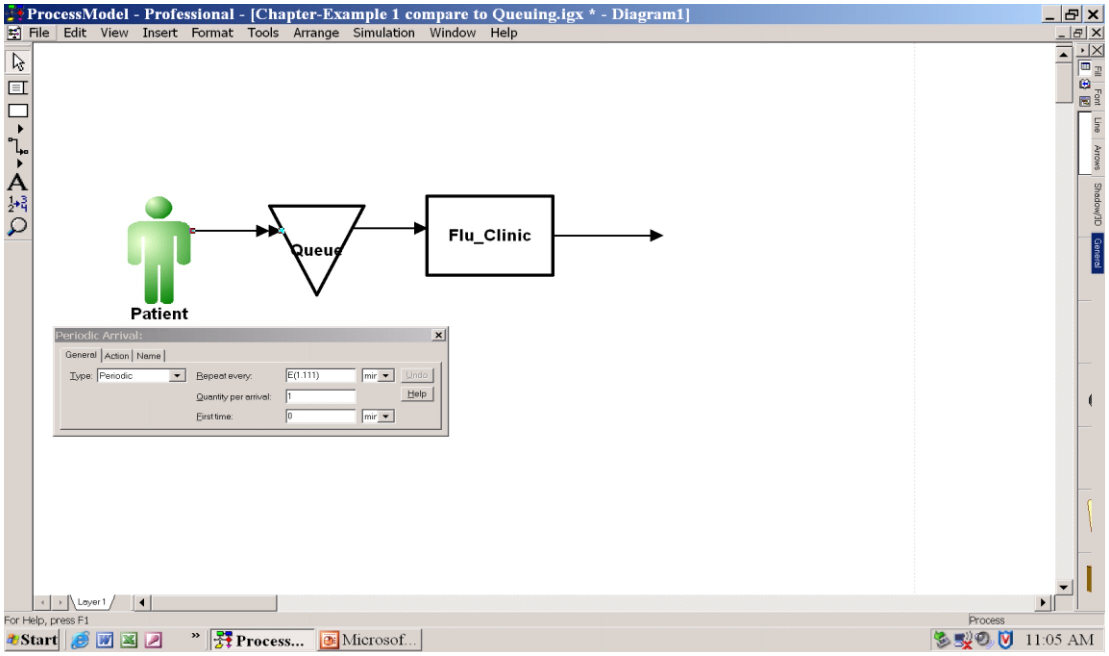
Queuing Analytic (QA) Application

In order to use QA we have to assume that patient arrival is the Poisson process with the average arrival rate λ=54 pts/hr, and shot time is exponentially distributed with the average time τ=2 min=0.033 hrs.

Using the number of servers *N*=2, and plugging the above data in the Excel spreadsheet with built-in QA formulas for the unlimited queue we get the average L_q =7.66 patients, and the average waiting time, *t*, about 8.5 min. The clinic's average utilization is 90%.

DES Application

DES model structure is presented on Fig 1.



Periodic Arrival:

General | Action | Name

Type: Periodic

Repeat every: E(1.111) mir

Quantity per arrival: 1

First time: 0 mir

Undo | Help

Fig 1. Layout of the simulation model of flu clinic. Information on the panel indicates patient arrival type (Periodic) that repeats on average every E(1.111) min (E stands for exponential distribution)

It simply depicts the arrived patient flow connected to Queue, then coming to the flu clinic (box called Flu_Clinic), and then exit the system. These basic model elements are simply dragged down from the pallet and then connected to each other.

Next step is to fill in the process information: patients arrive periodically, one patient at a random exponentially distributed time interval with the average inter-arrival time 60 min/54=1.111 min, E(1.111), as indicated on the data arrival panel on Fig 1. This corresponds to Poisson arrival rate 54 patients per hour (E stands for exponential distribution).

In the Flu_Clinic data panel the capacity input was 2 (two patients served concurrently by two nurses), and the service time was exponentially random with the average value 2 min, E(2). This completes the model set-up.

The model was run 300 replications to capture the exponential variability for different simulation time. Results are presented on Fig 2.

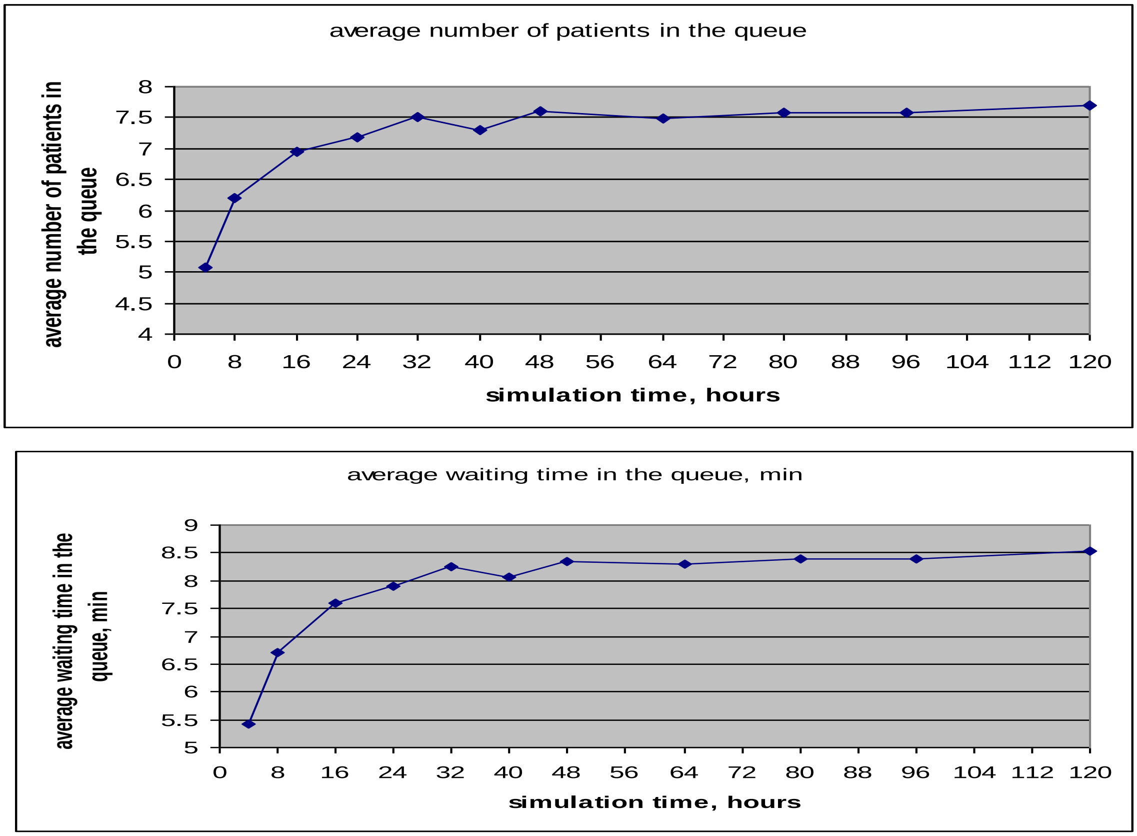


Fig. 2 *Unlimited queue size with steady-state operation.*
Average number of patients in the queue (top) and average waiting time in the queue (bottom).

It is seen that the number of patients in the queue steadily increases until a steady-state operation (plateau) is reached. The average steady-state number of patients in the queue is 7.6 with some small fluctuations around this average (top plot).

The average waiting time is presented on Fig 2 (bottom plot) with the steady-state value about 8.45 min with some variations around the average. The average steady-state utilization is 89.8%.

Thus, we received practically the same results with DES model as with QA model using about the same efforts.

Example 2

Flu clinic: unlimited queue size with non-steady-state operation

It was required to verify that the clinic would operate smoothly enough with a new team of two less experienced nurses who would work only a little slower than the previous one. The average time to make a shot will be about 2.5 min (instead of average 2 min for more experienced staff, as in the previous example).

It was reasoned that because this difference is relatively small it would not practically affect the clinic operation: the number of patients in the queue and their waiting time on the typical working day could be only a little higher than in the previous case or practically not different at all.

Queuing Analytic (QA) Application

The average service time 2.5 min = 0.0416 hrs and the same arrival rate 54 pts/hr were plugged in the *M/M/s* queuing calculator. However the calculator returned no number at all, which means that no solution could be calculated. Why is that ?

The average service time in this example is only slightly higher than it was in the previous example. However, this small difference made parameter $\rho = \lambda * \tau / N = 54*0.0416/2=1.125$ greater than 1. If this parameter is greater than 1 then a steady-state process does not exist. This explains why the calculations cannot be done using this value.

Queuing analytic formulas with unlimited queue size are applicable only for steady-state processes, i.e. for the established processes whose characteristics do not depend on time. The steady-state condition is possible only if $\rho < 1$, otherwise the queuing formulas are not applicable and the queue grows indefinitely.

It is demonstrated in the next section how DES methodology easily handles this situation and demonstrates growth of the queue.

DES Application

Using the same DES model layout described in example 1 we simply plug this average service time in the Flu_Clinic data panel making it E(2.5) min, and run the simulation model. Results are given on Fig 3.

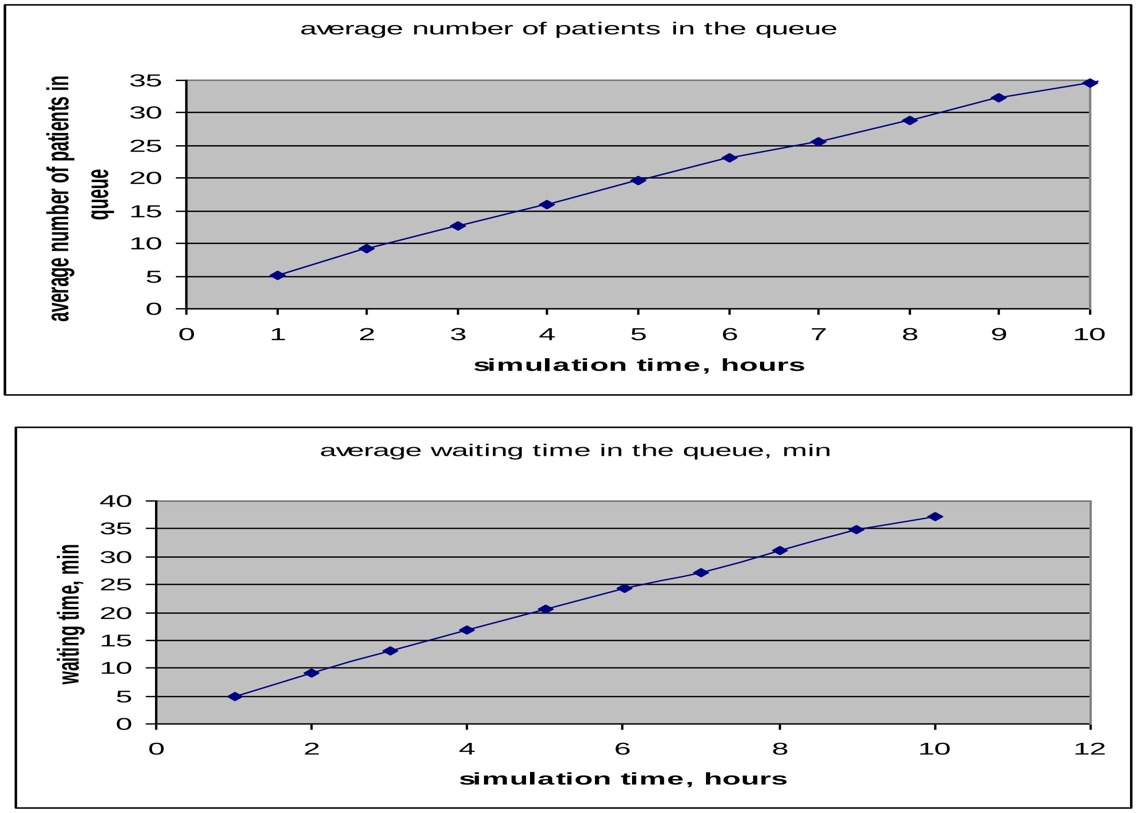


Fig 3 *Unlimited queue size with non-steady state operation.*
Average number of patients in the queue (top) and average waiting time in the queue (bottom).

These plots demonstrate how the patient queue (top plot) and waiting time (bottom plot) grow with clinic operation time. The plots demonstrate no apparent trend to a steady-state regime (plateau). The growth goes on indefinitely with time.

This example also illustrates an important principle of 'unintended consequences'. An intuition that is not supported by objective quantitative analysis says that a small change in the system input (service time from the average 2 min to 2.5 min) would result in a small change in the output (small increase in the number of waiting patients and their waiting time). For some systems it is indeed true. Systems in which the output is always directly proportional to input are called linear systems. However there are quite a few systems in which this simple reasoning breaks down: a small change in the value of system's input parameter(s) results in a dramatic change (even qualitative change) in the system's outcome (behavior), e.g. from a steady-state regime to a non-steady-state regime. Such systems are called non-linear or complex systems despite the fact that they can consist of only a few elements.

Example 3
Limited queue size with 'inpatient' patients leaving the system

Unlimited queue size is not always a good model of real systems. In many cases patients wait in a waiting lounge that has usually a limited number of chairs (space). QA models designated *M/M/s/K* are available that include a limited queue size, *K* (Green, 2006; Lawrence and Pasternak, 1998; Hall, 1990). However the analytic formulas become very cumbersome. If the QA model includes some patients that leave the system after waiting some time in the queue, the formulas become almost intractable.

In contrast, DES models easily handle the limited queue size and patients leaving before the service began (‘inpatient’ patients).

To illustrate, we use the same DES model as in Example 1 with only a slight modification to include a new limited queue size and ‘inpatient’ patients.

Suppose that the queue size limit is 10 (chairs or beds) and patients leave after waiting 10 min (of course, these could be any numbers including statistical distributions). We put 10 in the field ‘Input queue size’, and draw a routing ‘renege after 10 min’. The new model is now ready to go. Simulation results are presented on Fig 4 and Fig 5.

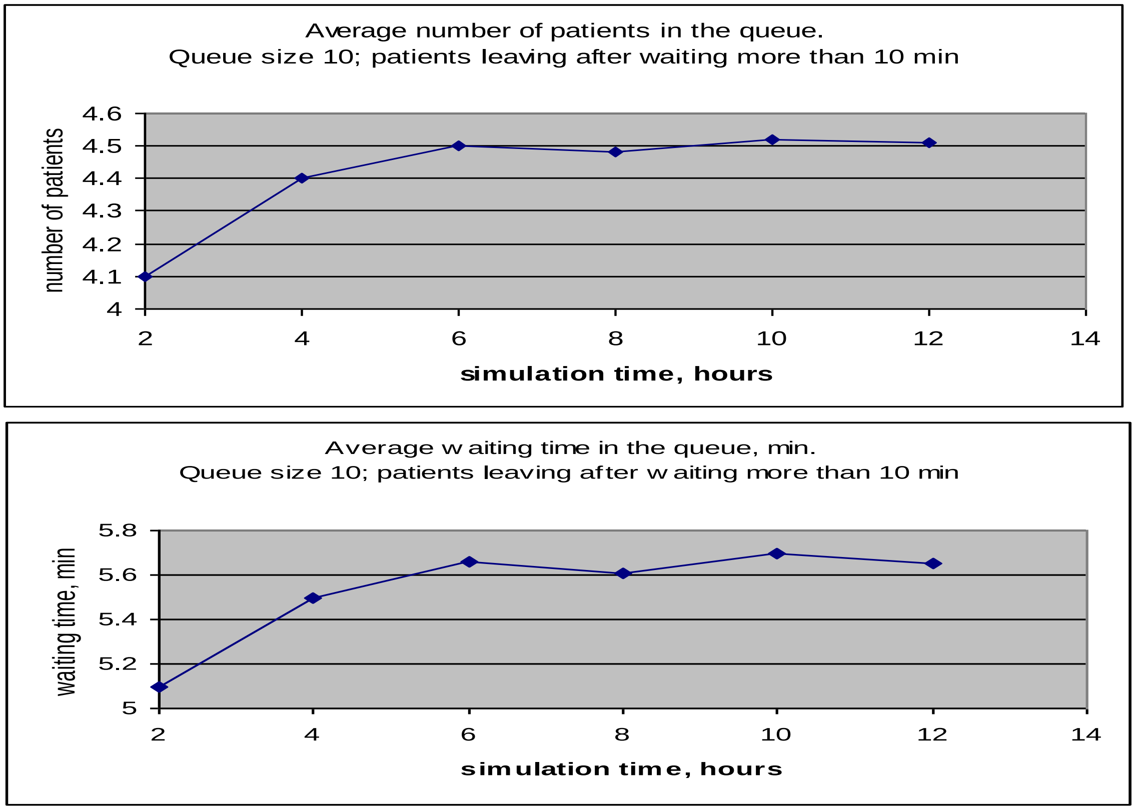


Fig 4 *Limited queue size with ‘inpatient’ patients leaving the system.*
Average number of patients in the queue (top) and average waiting time in the queue (bottom).

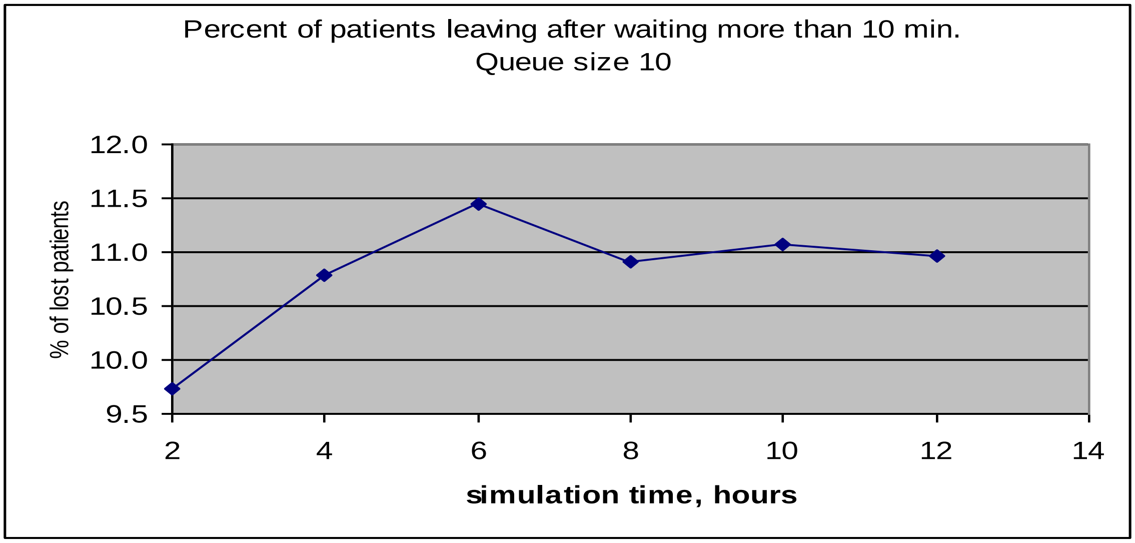


Fig 5 *Limited queue size with 'inpatient' patients leaving the system.*
Percent of patients leaving the system (reneging) after waiting more than 10 min. Input queue size 10.

The difference between an unlimited queue size (*Example 2*) and a limited one with leaving patients is significant.

The plots suggest that limited queue size results in a steady-state solution (plateau). (It could be proved analytically that a steady-state solution always exists if the queue size is limited). However 10% to 11 % of patients are lost because they did not want to stay in the queue more than 10 min (Fig 5).

Thus, this simple DES model gives a lot of valuable information and serves as powerful tool to find out how to better manage the flu-clinic.

Example 4

Flu clinic: time-varying arrival rates

In the previous examples parameters of the queuing system (average arrival rate 54 patients per hour and average shot time 2.5 min) made a patient flow a non-steady-state one for which QA model with unlimited queue size could not be used.

However, the average patient arrival rate varies significantly during a day, and 54 patients per hour was actually a peak arrival rate, from noon to 3 pm. In the morning hours from 8 am to 10 am the arrival rate was lower, 30 patients / hour. From 10 am to noon it was 40 patients / hour, and in the afternoon from 3 pm to 6 pm it was about 45 patients / hour.

Thus, the average arrival rate for these time periods for the day was calculated as (30+40+54+45)/4=42.25 patients/hour.

Queuing Analytic (QA) Application

The daily average arrival rate 42.25 was plugged in the queuing calculator (along with the average time to make a shot 2.5 min). The calculator returned the average number of patients in queue L_q =1.4 patients, and the average waiting time about 1.97 min. Because the calculator produced some numbers it was concluded that this clinic process will be in a steady-state condition and that the waiting time and the number of patients in the queue is acceptable.

But was it a correct conclusion ? Recall that QA models assume that a Poisson arrival rate is constant during a steady-state time period (Hall, 1990; Lawrence et al, 2002; Green, 2006). If it is not constant, such as in this case, QA results could be very misleading. The wait time will be significantly greater in the mid-day period (and/or the steady-state condition will be violated). At the beginning and at the end of the day, though, the wait time will be much smaller.

Because the arrival rate is included non-linearly in the exponential term of a Poisson distribution formula, the arrival rate cannot be first averaged and then substituted in the exponential term. (For non-linear functions, the average value of a function is not equal to the function of average values of its arguments).

As Green (2006) stated '…this illustrates a situation in which a steady-state queuing model is inappropriate for estimating the magnitude and timing of delays, and for which a simulation model will be far more accurate'.

It is tempting, as a last resort, to save the use QA approach by dividing the day into time periods in which arrival rate is approximately constant. Then a series of *M/M/s* models is constructed, one for each period. This approach is called SIPP (stationary independent period-by-period) (Green, 2006; Green et al, 1991).

If we apply this approach, the following results can be obtained:

Time period 8 am to 10 am: L_q =0.8 patients in the queue, waiting time 1.6 min
Time period 10 am to noon: L_q =3.8 patients in the queue, waiting time 5.7 min
Time period noon to 3 pm: *no steady-state solution*
Time period 3 pm to 6 pm: L_q =13.6 patients in the queue, waiting time 18.1 min

Notice how these results differ from those based on the averaging of the arrival rate for the entire day.

However this SIPP patch applied to QA models was found to be unreliable (Green, 2006; Green et al, 2001). This is because in many systems with time-varying arrival rates, the time of peak congestion significantly lags the time of the peak in the arrival rate (Green et al, 1991). These authors developed a modification called *Lag-SIPP* that incorporates an estimation of this lag. This approach has been shown to be often more effective than a simple SIPP (Green, 2006).

Even it is so, this does not make QA models application easier if there are many time periods with different constant arrival rates because many different *M/M/s* models need to be constructed accordingly to describe one process.

Let's now see how DES approach easily and elegantly handles this situation with time-varying arrival rate.

DES Application

The DES model structure (layout) for time-varying arrival rate is the same as it was used in Example 1 (Fig. 1).

The only difference will be a different arrival routing type: instead of periodic arrival with the random inter-arrival time, an input daily-pattern arrival panel should be used. We use one day of the week, and input 60 patients from 8 am to 10 am (30 pts/hour *2); 80 patients from 10 to noon (40 pts/hour*2); 162 patients from noon to 3 pm (54 pts/hour*3); and 135 patients from 3 pm to 6 pm (45 pts/hour * 3). The model of the entire day (from 8 am to 6 pm) is ready to go.

The following simulation results are obtained (compare with the approximated QA SIPP model results from above):

Time period 8 am to 10 am: L_q =0.6 patients in the queue, waiting time 0.84 min
Time period 10 am to noon: L_q =2.26 patients in the queue, waiting time 2.75 min
Time period noon to 3 pm: L_q =14.5 patients in the queue, waiting time 15.3 min
Time period 3 pm to 6 pm: L_q =20 patients in the queue, waiting time 25.5 min

It is seen that QA SIPP model over-estimates the queue at the beginning of the day, under-estimates the queue at the end of the day, and provides no results at all for the middle of the day (noon to 3 pm).

We have to conclude that QA approach should not be used for time-varying arrival rates.

Example 5

ICU waiting time

This problem is presented by (Litvak 2007; Weber, 2006) to demonstrate how QA can be used to calculate patient average steady-state waiting time to get into ICU if it has 5 beds and 10 beds and patient arrival rate is 1 per day and 2 per day, accordingly. The author assumes that the average length of stay in the ICU is 2.5 days and, of course, that patient arrival is a Poisson process.

In order to apply QA formulas an additional assumption should be used that length of stay follows exponential distribution with the above average value.

QA Application

Using *M/M/s* model it is easy to calculate that the average waiting time for 10 beds ICU is 0.43 hours, and that for 5 beds ICU is 3.1 hours. Average ICU utilization is 50%.

Thus, the waiting time for the larger unit is about 7 times shorter (the author calls it tenfold shorter for the larger unit, Litvak, 2007).

Notice, however, that this result is valid only for exponential service time. If we use other more realistic distributions for length of stay with the same average we should get a different result.

For example, length of stay could be in the range from 2 to 3 days with the average 2.5 days, and be described by a triangle distribution. Or length of stay could follow a long-tailed log-normal distribution, also with the same average 2.5 days and standard deviation, say, 2 days (these values would correspond to log-normal parameters 3.85 and 0.703).

QA does not distinguish between different distributions, and always produces the same result if the same average is used regardless of the actual distribution. This is a serious limitation of QA.

Let’s demonstrate how easy to use DES for different length of stay distributions.

DES application

We can use exactly the same model as in the previous examples (Fig. 1). We simply plug capacity 5 or 10, accordingly. We start with exponential distribution with the average 2.5 days, E(2.5), for length of stay to compare with QA.

We also plug the average inter-arrival time as 1 day or 0.5 days, accordingly. Steady-state DES waiting times are: 0.43 hours for 10 beds ICU and 2.94 for 5 beds ICU, accordingly. These are practically the same results as for QA.

Now, let’s see how different distributions with the same average length of stay affect the waiting time. Recall, that QA cannot at all answer such practically important questions, and valid only for exponential distributions (or, at best, for distributions with coefficient of variation close to1, Green, 2006).

For triangle distribution limited between 2 days and 3 days with the average 2.5 days, we get:
for 10 beds ICU average waiting time is 0.27 hours while for 5 beds unit it is 1.72 hours. Notice, how significantly differ these values from the exponential length of stay with the same value.

Similarly, for log-normal distribution with the same average 2.5 days and standard deviation 2 days, we get:
for 10 beds ICU average waiting time is 0.35 hours while for 5 beds unit it is 2.46 hours.

Thus, QA is severely limited in what it cannot take into account different distributions of service time and always produces the same result if the same average is used regardless of the different distributions with the same average.

Example 6

***Mixed patient arrivals: random and scheduled*.**

We frequently deal with mixed patient arrivals pattern, i.e. some patients are scheduled to arrive on specific time while some patients arrive unexpectedly. For example, some clinics accept patients who made an appointment and also accept urgent random walk-in patients. Operating room suites schedule elective surgeries while suddenly a trauma patient arrives and an emergency surgery is required. Such a mixed arrival pattern with a different degree of the variability requires a special treatment.

QA models should not be used if arrival flow contains a non-random component, i.e. it is not a Poisson random.

Let’s illustrate what happens if this principle is violated.

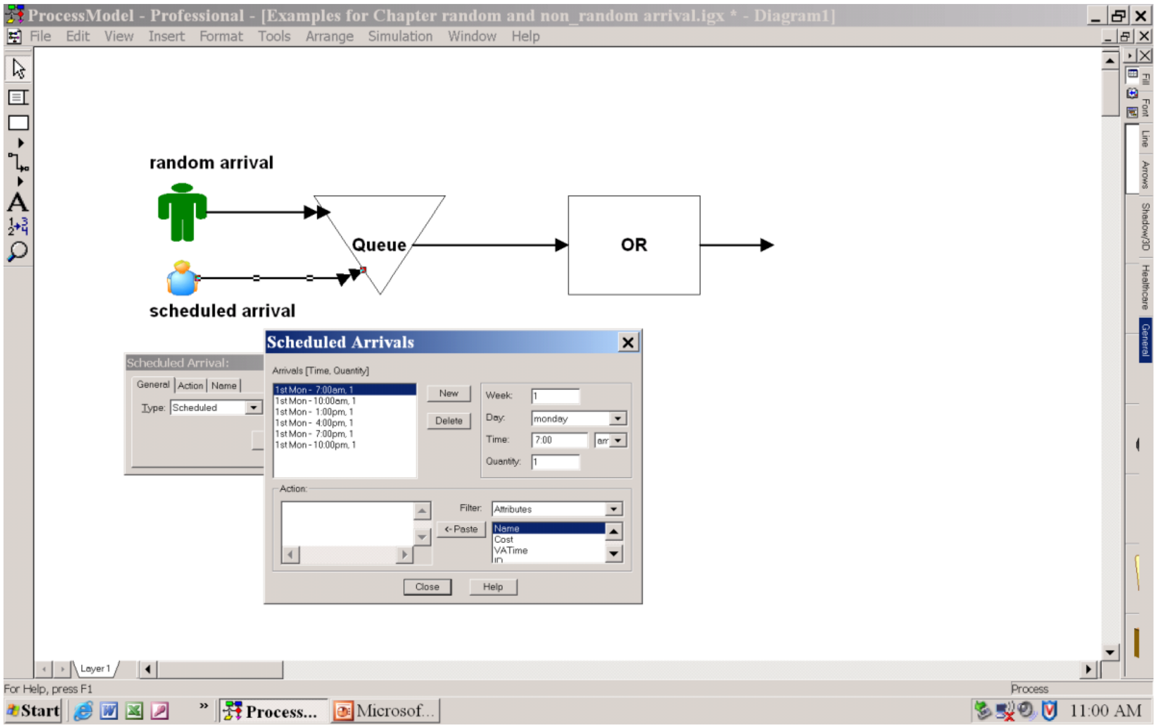
Suppose that there is one operating room (OR) and there are six scheduled surgeries for a day, at 7 am, 10 am, 1 pm, 4 pm, 7 pm, 10 pm. On this day six random emergency patients also arrived with the average inter-arrival time 4 hours, i.e. E(4) hours. Total number of patients for one day is 12.

Queuing Analytic (QA) application

If QA model is applied assuming that all 12 patients are random arrival then we would get the average arrival rate 12 pts/24= 0.5 pts per hour. Using the average surgery time 1 hour, E(1) hours, we get the average number of patients in the queue, L_q = 0.5, waiting time in queue, W_q=1 hour, and time in the system, W_s =2 hours.

DES Application

We use a simple DES model with two arrival flows, one random, E(4) hours, and another one with scheduled six patients, as indicated on Fig 6.



Type: Scheduled

Arrivals [Time, Quantity]

1st Mon - 7:00am, 1
1st Mon - 10:00am, 1
1st Mon - 1:00pm, 1
1st Mon - 4:00pm, 1
1st Mon - 7:00pm, 1
1st Mon - 10:00pm, 1

Week: 1
Day: monday
Time: 7:00 am
Quantity: 1

Filter: Attributes

Name
Cost
VATime
ID

Fig 6 Mixed patient arrivals: random and scheduled. Two arrival flows, one random, E(4) hours, and one with six patients scheduled at 7 am, 10 am, 1 pm, 4 pm, 7 pm, 10 pm, as indicated on the panel.

Running simulation for 24 hours, we get: the average number of patients in the queue, L_q = 0.3, waiting time in queue, W_q=0.33 hours, time in the system, W_s =0.55 hours.

Notice how badly QA model over-estimated the time: almost by a factor of 3 for waiting time in the queue, and almost by a factor of 4 for time in the system !

QA models cannot account accurately enough for arrival variability that is lower than Poisson variability.

There are some approximate QA formulas that include a coefficient of variation of arrival and service time distribution but only for one server (Green, 2006).

This illustrates a general principle: the lower variability in the system (both arrival and service) the lower delays (see also Green, 2006).

In other words, lowering variability is the key to improving patient flow and to reduce delays and waiting times.

One of the root causes of why intuition usually fails to account for the effect of variability even in very simple systems is because one usually operates with the average values but actually treats them as fixed values with no variability information. DES models, though, naturally handle variability using statistical distributions with multiple replications.

Example 7

Effect of added variability on process flow and delay

Let's now demonstrate how additional arrival variability, service time variability, and/or both of them would affect the throughput and waiting time in the system. We will be using a simple DES model similar to the model presented on Fig 1.

We consider five scenarios with consecutively added step-by-step patient flow variability:

Scenario 1. No variability at all. Each patient arrives exactly every 2 hours. Service time (surgery) is exactly 2 hours.

Scenario 2. There is arrival variability, i.e. a Poisson flow with the average arrival rate 0.5 pts/hr (average inter-arrival time is 2 hours, E(2) hrs). No service time variability, it is exactly 2 hours.

Scenario 3. No arrival variability. There is a service time variability with the average service time 2 hours, E(2) hours.

Scenario 4. There are both types of variability. Poisson arrival with the average arrival rate 0.5 patients per hour (average inter-arrival time 2 hours), and service time variability with the average service time 2 hours, E(2)

Scenario 5. Poisson arrival variability with the average arrival rate 0.5 patients per hour (average inter-arrival time 2 hours). Service time variability is Log-normally distributed with the distribution mean value 2 hours and standard deviation 4 hours (these values correspond to the Log-normal parameters: location= -0.112 and scale=1.27).

Results for 24 hours simulation time are summarized in Table 1.

Notice that QA could only be applied for Scenario 4.

It follows from this table that as the variability steps are added to the process, patient throughput decreases, an overall waiting time increases, and utilization decreases. At the same time variability should not be characterized only by a single parameter, such as its coefficient of variation (CV). The overall shape of the variability distribution also plays a role. For example, coefficient of variation for the log-normal distribution service time (CV=4/2=2) is greater than that for exponential distribution (CV=1). Nonetheless, this did not result in increase of the waiting time, as this would follow from an approximated queuing formula (Allen, 1978; Green, 2006). Only DES modeling can accurately account for the effect of the distribution shape and skewness.

Conclusions

Examples presented here illustrate that Queuing Analytic (QA) spreadsheets have rather limited applications. Practically QA could be reliably used only in Example 1 which represents the simplest case. Even small complications, such as a non-steady state flow, limited queue size with leaving patients, time-varying arrival rate, non-random arrival or non-exponential case duration, make QA applications difficult, inaccurate or impossible at all.

On the other hand, DES models are flexible, give more information than QA, and it is easy to build them using convenient software user interface.

Certainly, building a complex simulation model of hospital or unit operations requires efforts for custom logic development, debugging, model validation, and input data collection. Such a simulation goes well beyond examples presented here and it is presented elsewhere. But these examples clearly illustrate why in most cases DES is superior and preferred to QA.

It should be noted that Institute for Healthcare Improvement (IHI) regularly promotes through its workshops and seminars the use of Queuing Theory as the main quantitative methodology to address pressing hospital problems of patient flow and variability (IHI, 2008; Litvak, 2007; McManus et al, 2004; Haraden et al, 2003).

In contrast to picking the right queuing model (if it available at all), DES model is custom built using standard modeling elements, their connections and action logic that reflect the structure of the concrete and often unique system of interest. This gives DES approach a real flexibility and advantage over un-flexible QA.

Unfortunately, a non-critical application of queuing theory brings more harm and confusion than good to most healthcare administrators who usually do not have enough experience and background in operations research. As a result, analytic queuing theory application is often misused, misplaced and misleading, as highlighted, for example, by D'Alesandro, J., 2008.

Hopefully, material presented here will help to look beyond analytic queuing theory, and turn over to DES modeling as a methodology of choice.

Table 1 Five DES scenarios with consecutively added variability. Simulation performed for 24 hours period.

INPUT		OUTPUT FOR ONE DAY (24 hours)				
Patient Arrival	Time in OR	average throughput (number of patients)	average number of patients waiting in queue	average time in the system, hrs	average waiting time in the system, hrs	average OR utilization,%
one patient every 2 hrs (no variability)	2 hrs (no variability)	12	0	2	0	100%
arrival variability: average inter-arrival time 2 hrs (Poisson arrival rate 0.5 pts/hr)	2 hrs (no variability)	10	1.4	4.1	2.1	86%
one patient every 2 hrs (no variability)	service time variability: average 2 hrs (exponential time distribution)	10	1	3.5	1.6	82%
arrival variability: average inter-arrival time 2 hrs (Poisson arrival rate 0.5 pts/hr)	service time variability: average 2 hrs (exponential time distribution)	9.4	1.6	4.2	2.3	78%
arrival variability: average inter-arrival time 2 hrs (Poisson arrival rate 0.5 pts/hr)	service time variability: average 2 hrs, standard deviation 4 hours (Log-normal distribution parameters: loc= -0.112; scale=1.27)	9.3	1.4	3.3	1.7	69%

REFERENCES

Abu-Taieh, E., El Sheikh, A R., 2007. *Commercial Simulation Packages: A Comparative Study*. International Journal of Simulation, vol 8, No 2, pp. 66-76

Allen, A., 1978. Probability, statistics and queuing theory, with computer science applications. New York, Academic Press

D'Alesandro, J., 2008. *Queuing Theory Misplaced in Hospitals*. Management News from the Front. Process Improvement. PHLO. http://phlo.typepad.com. Posted Feb 19, 2008

Gallivan, S., Utley, M., Treasure, T., Valencia, O., 2002. *Booked inpatient admissions and hospital capacity: mathematical modeling study*. British Medical Journal, Feb 2, 324: 280-282

Glantz, S., 2005. Primer of Biostatistics. 6-th ed. New York, McGraw-Hill Co., Inc .

Green, L. 2006. *Queuing Analysis in Healthcare*. In Hall, R. (Ed.), *Patient Flow: Reducing Delay in Healthcare Delivery (pp. 281-307). Springer, NY*

Green, L., 2004. *Capacity Planning and Management in Hospitals*. In Brandeau., M., Sainfort, F., Pierskala, W., (Eds.). Operations Research and Health Care. A Handbook of

Methods and Applications.(pp.15-41). Kluwer Academic Publisher, Boston/Dordrecht/London

Green., L., Kolesar, P., Svoronos, A., 1991. Some effects of non-stationarity on multi-server Markovian queuing Systems. Operations Research, 39, p.502-511

Green, L, Kolesar, P., Soares, J. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research, 49, p.549-564

Hall, R. 1990. Queuing methods for Service and Manufacturing. New Jersey, Prentice Hall

Haraden, C., Nolan, T., Litvak, E., 2003. *Optimizing Patient Flow: Moving Patients Smoothly Through Acute Care Setting.* Institute for Healthcare Improvement Innovation Series 2003. White papers 2, Cambridge, MA

Harrison, G., Shafer, A., Mackay, M., 2005. *Modeling Variability in Hospital Bed Occupancy.* Health Care Management Science, 8, p. 325-334

Hillier, F., Yu, O., 1981. Queuing Tables and Graphs. New-York, Elsevier, pp.1-231

Hlupic, V., 2000. *Simulation Software: A Survey of Academic and Industrial Users*. International Journal of Simulation. Vol 1, No 1, pp.1-11

(IHI)- Institute for Healthcare Improvement, 2008. Boston, MA http://www.ihi.org/IHI/Programs/ConferencesAndSeminars/ApplyingQueuingTheorytoHealthCareJune2008.htm

Jacobson, H., Hall, S., Swisher, J., 2006. Discreet-Event Simulation of Health Care Systems. In Hall, R. (Ed.), *Patient Flow: Reducing Delay in Healthcare Delivery (pp. 210-252). Springer, NY*

Kopach-Konrad, R., Lawley, M., Criswell, M., Hasan, I., Chakraborty, S., Pekny, J., Doebbeling, B., 2007. *Applying Systems Engineering Principles in Improving Health Care Delivery.* Journal of General Internal Medicine, 22 (suppl 3): 431-437

Lawrence, J., Pasternak, B., Pasternak, B.A. 2002. Applied Management Science: Modeling, Spreadsheet Analysis, and Communication for Decision Making. John Wiley & Sons.

Litvak, E., 2007. A new Rx for crowded hospitals: Math. Operation management expert brings queuing theory to health care. American College of Physicians-Internal Medicine-Doctors for Adults, December ACP Hospitalist

Litvak, E., Long, M., 2000. *Cost and Quality under managed care: Irreconcilable Differences?* Am. Journal of Managed Care. 6 (3). p. 305-312

Mayhew, L., Smith, D. 2008. *Using queuing theory to analyze the Government's 4-h completion time target in Accident and Emergency departments*. Health Care Management Science, 11, p.11-21.

McManus, M., Long, M., Cooper, A., Litvak, E., 2004. *Queuing Theory Accurately Models the Need for Critical Care Resources.* Anesthesiology, 100(5), p. 1271-1276,

McManus, M., Long, M., Cooper, A., Mandell, J., Berwick, D., Pagano, M., Litvak, E., 2003. *Variability in Surgical Caseload and Access to Intensive Care Services***.** Anesthesiology, 98(6), p. 1491-1496

Nikoukaran J., 1999. *Software selection for simulation in manufacturing: A review*. Simulation Practice and Theory. Vol 7, No 1, pp.1-14

Seelen, L., Tijms, H., Van Hoorn, M., 1985. Tables for multi-server queues. New-York, Elsevier, pp. 1-449

Swain, J., 2007. *Biennial Survey of discreet-event simulation software tools*. OR/MS Today, v.34, N 5, October, The Institute for Operations Research and the Management Science. Lionheart Publishing, Inc. http://www.lionhrtpub.com

Weber, D. O., 2006. Queue Fever: Part 1 and Part 2. *Hospitals & Health Networks*, Health Forum. May, 2006. http://www.IHI.org