5-2010

Applications of Queueing Models in Hospitals

Foad Mahdavi Pajouh
Oklahoma State University, mahdavi@okstate.edu

Manjunath Kamath
Oklahoma State University, m.kamath@okstate.edu

Applications of Queueing Models in Hospitals

Foad Mahdavi Pajouh
Oklahoma State University
mahdavi@okstate.edu

Manjunath Kamath
Oklahoma State University
m.kamath@okstate.edu

ABSTRACT

In a hospital system, increasing resource utilization to reduce costs and decreasing patients' waiting time to provide timely care and improve patient satisfaction are important but conflicting goals. Queueing models can provide reasonably accurate evaluations of system performance and are popular among researchers and system designers because of their analytical nature and their ability to provide quick solutions for "what-if" analyses. There is a considerable amount of published research on using queueing to analyze and design hospital facilities. We review and categorize this literature in an attempt to motivate further research in applying queueing models in the healthcare domain.

Keywords

Queueing theory, Healthcare, Hospital, Waiting time.

INTRODUCTION

Queueing models have been extensively used to model and analyze different healthcare systems such as hospitals (Cochran and Roche, 2009), pharmaceutical industry (Viswanadham and Narahari, 2001), and organ transplant (Zenios, 1999). Nearly a decade ago, Preater (2001) compiled a bibliography of queueing applications in healthcare. This paper reviews and categorizes applications of queueing models in hospital systems. Based on the nature of problems, the two major areas of application are (i) design and analysis and (ii) planning and scheduling. In the first group, the focus is more on the emergency services. The emergency services studied are further sub-divided into (a) emergency department and (b) ambulance system. In the second group (planning and scheduling), the focus is on (a) staffing and (b) patient planning problems. Figure 1 shows the aforementioned taxonomy. Standard notation for describing queueing models has been used in the following sections and the reader is referred to Gross and Harris (1998) for more details.

Design and Analysis

Emergency Department

Inadequate or delayed service in the emergency department may result in a huge cost to the system and sometimes result in death or serious injuries to patients. On the other hand, using expensive resources (doctors and medical equipment) for situations other than those intended (true emergencies) in this department may decrease the effective utilization of these resources and increase the service cost. So it is important for hospital managers to model emergency department's operations and analyze its performance in order to manage the overall costs. Another interesting characteristic of the emergency department is that patients arrive randomly which makes modeling arrivals in this department a challenging task.

Solberg et al. (2003) listed 38 measures for evaluating emergency department performance. One the most important measures mentioned in this work is *leaving without treatment ratio* (LWTR), which has been studied by several other researchers. A critical factor that increases LWTR is patients' long waiting times to start their service. Queueing models provide efficient and useful models to establish decision support systems for minimizing patients' waiting time and maximizing emergency department's performance. Broyles and Cochran (2007) modeled the emergency department as a production network by using the $M/M/1/K$ model. They also claimed that besides waiting time, the patients' points of view toward delay may also affect LWTR. In this work, the authors estimated LWTR by using a nonlinear regression on the "full queue" formula of $M/M/1/K$ model. Then they used this LWTR to estimate emergency department's business loss. They suggested that redesigning the system, e.g., separating patients by their type of service may help reduce LWTR. Roche and Cochran (2007) used queueing networks to model patient flow in an extreme "fast-track" emergency department design. The authors first determined peak and off-peak arrival rates to each node in the network and then they used queueing equations to estimate the size of the waiting area in the emergency department. They found that separating non-urgent patients and treating them in fast-track areas reduces LWTR. Cochran and Roche (2009) used an open queueing network model to design an emergency department in order to increase its capacity. The authors used waiting time and overflow probability as system performance indicators in this work. They concluded that population growth, unavailability of emergency departments nearby and variation of seasonal peak can considerably increase patients' average waiting time, which leads to an increase in LWTR.

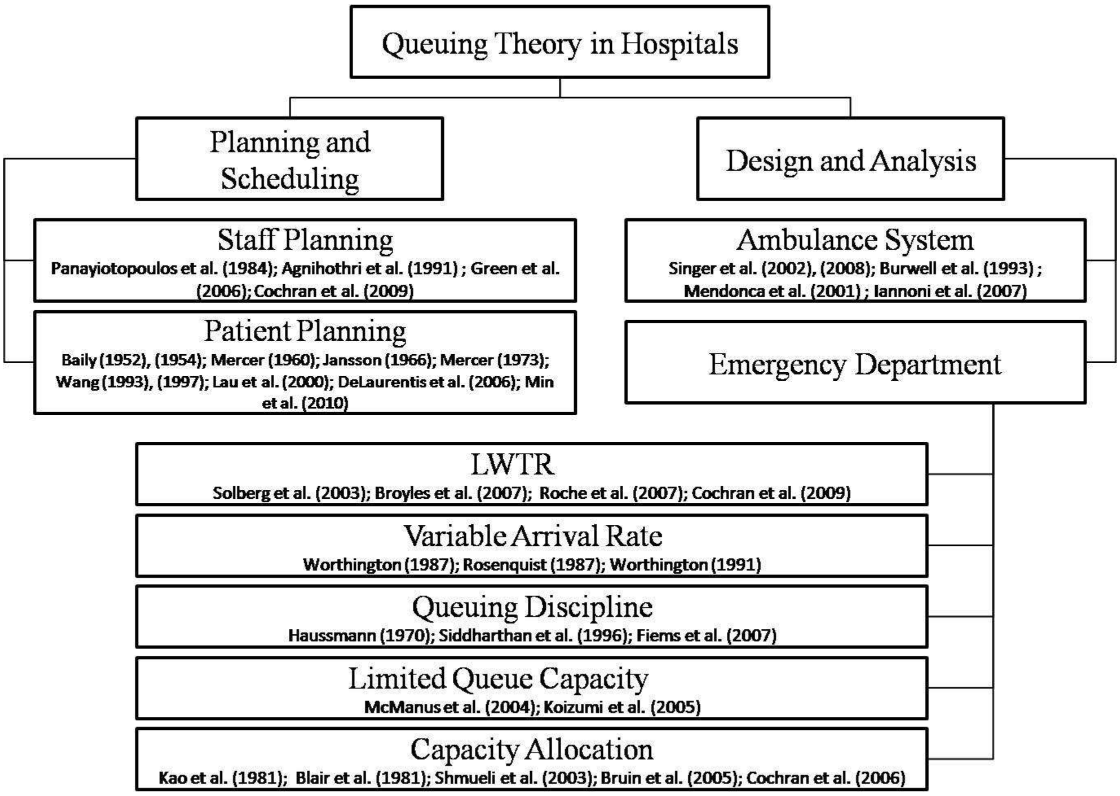

Figure 1. A Taxonomy of Literature on Applications of Queueing Theory in Hospitals

Some researchers have considered variable arrival rates in modeling the emergency department. Worthington (1987) estimated the effect of management actions on waiting lines. Their model assumes that service times are sampled independently from a fixed probability distribution and arrivals occur randomly at a rate that decreases linearly with waiting-list size. The author claimed that "feedback" is an effective factor in waiting-list management and tightening up the available feedback system will provide an effective tool for controlling the waiting list. Rosenquist (1987) determined the characteristics of service and inter-arrival times for an emergency room radiology service. To demonstrate the usefulness of queueing analysis in radiology, the author provided two examples. In the first, the researcher studied the effect of 5% annual increase in demand on patients' waiting time and in the second one, queueing models were used for cost analysis. Worthington (1991) did some work with hospital consultants in Lancaster District Health Authority to develop and utilize waiting list management models. The author found that the traditional method for reducing waiting time, i.e., increasing service capacity, is not an effective approach since the arrival rate of patients increases as the patients realize that service times may decrease which again increases the waiting time.

Queueing discipline is an important factor that may significantly effect waiting time for service. Haussmann (1970) used queueing theory to develop a model of the patient care process for application to a burn unit and used this model to evaluate the expected waiting times of various priorities of patients. These waiting times were used as a quality of service (QoS) indicator to study the effect of changes in service or demand rates. The relationships among three factors: patient condition, nurses' activity priorities and patient load per nurse were also studied in this paper. Siddharthan et al. (1996) studied the inappropriate use of the emergency department by patients seeking non-emergency or primary care and used an economic model to show that this may increase the waiting time in an emergency department. The authors proposed a priority queueing model to reduce average waiting times. The priority of a patient was determined based on the criticality of the required service. The waiting time of higher priority patients decreases while lower priority patients have more waiting time on the average. Fiems et al. (2007) investigated the impact of emergency requests on the waiting times of scheduled patients. It was also assumed that an emergency request interrupts the scheduled patient's examination and the scheduled patient's service has to start all over. The authors used a discrete-time queueing model for their study.

Some studies considered the situation where the queue length has a limit. When the queue reaches its maximum length allowed, walk-in patients will be turned away until the queue length decreases. This event is called "blocking." McManus et al. (2004) constructed a mathematical model of patient flow in a busy urban intensive care unit. The proposed queueing model could predict admission turn-away (blocking) accurately with correlation coefficient of 0.89. Using the proposed queueing model, they also found that the system performance would drastically decrease with even a small change in servers' availability. Koizumi et al. (2005) used a queueing network system with blocking to analyze the healthcare system congestion processes. It was assumed that arrival rates and number of servers are constant in the model. The authors found that one of the main reasons for system congestion is existence of facility-specific bottlenecks and removing such bottlenecks may efficiently reduce congestion in the system.

Capacity allocation to different areas of an emergency department is an important problem. Researchers consider available beds in a unit as that unit's capacity and try to find the optimum number of beds in each unit. Kao and Tung (1981) proposed an approach based on queueing models to periodically reallocate beds to services to minimize the expected overflows. Queueing models were used to estimate the patient population dynamics for each service. They showed that queueing based models are simple and useful tools for analyzing the bed allocation problem in a healthcare facility. Blair and Lawrence (1981) used finite capacity multi-server queues and continuous-time Markov chains to model a system of burn care facilities linked together by a referral policy to accommodate patient overflow. They also used a heuristic optimization procedure to find the optimal setting in a burn care facility in New York State. The goal was to maintain a 95% level of service and keep LWTR less than 5%. They found that their proposed approach is ideal for a system with low demand and high infrastructural costs. Shmueli et al. (2003) proposed a model to maximize the expected incremental number of lives saved from operating an intensive care unit (ICU). They used single-queue models to find the probability distribution of the number of occupied ICU beds. They modeled the ICU at Jerusalem's Hebrew University-Hadassah Hospital by using the proposed methodology and showed that a relative life saving improvement of 17.9% could be achieved by reforming the ICU admission policy. Bruin et al. (2005) applied a stationary 2-D queueing system with blocking to analyze congestion in emergency care chains. The authors' primary goal was to determine the optimal bed allocation over the emergency care chain, given a required service level (max. 5% refused admissions). They successfully identified bottlenecks, described the impact of fluctuation in demand and calculated the optimal bed capacity distribution. Cochran and Bharti (2006) used queueing network analysis and optimization to balance inpatient bed unit utilizations in a hospital and analyzed a 400 plus bed major hospital.

Ambulance System

In case of a medical emergency, it is very important for ambulance services to reach the site as fast as possible. Patient waiting time in this situation is a key indicator for ambulance system performance. Singer et al. (2002) studied the problem of a fleet configuration in which emergency vehicles receive demand calls while they are on the road. Their goal was to minimize the operational cost subject to constraints including maximum vehicle size and maximum waiting time of a patient. They showed that by sharing the buffer of orders between a set of vehicles, the customer waiting time can be reduced up to 10%. Singer and Donoso (2008) also used queueing theory to support decision-making in the ambulance business in Chile and calculated managers' and patients' key performance indicators separately. They used the proposed model to evaluate the impact of operational enhancements and optimize the geographical coverage of the bases.

The hypercube model which is a spatially distributed queueing model based on Markovian approximations has been one of the most popular techniques to model emergency vehicle systems (Burwell et al., 1992; Larson, 1974 and 1975). Since the hypercube model was not originally introduced to model emergency systems, some critical characteristics of such systems (e.g. dispatch ties) were not addressed in this model. Burwell et al. (1993) developed an extension of the hypercube model that contains preference ties and applied the proposed model to the emergency medical system of Greenville County, SC. They concluded that the proposed model could provide good estimates of the emergency system performance when input parameters are accurately specified. Mendonca and Morabito (2001) used the hypercube model to evaluate the mean response time of the system to an emergency call in an emergency medical system on a Brazilian highway connecting the cities of Sao Paulo and Rio de Janeiro. They showed that the proposed model can provide a powerful decision support tool which reduces the workload unbalancing among the ambulances. Using this model, they were also able to reduce the percentage of the lost calls to 5% of the total received calls. Iannoni and Morabito (2007) used the hypercube model to analyze emergency medical systems, which can receive calls on-line on highways. They considered partial backup, multiple dispatch, different classes of servers and customers and particular dispatching policies in the model. The proposed model can be directly embedded into optimization procedures for ambulance deployment to maximize the expected coverage or determine optimal primary response areas based on the ambulances' locations.

Planning and Scheduling

Staff Planning

Staff planning in order to satisfy the demand is one of the areas in which queueing models can be used effectively. Queueing models can help planners to estimate the number of required staff in each unit to achieve an acceptable customer LWTR. Panayiotopoulos and Vassilacopoulos (1984) developed a methodology based on $GI/G/C(t)$ model to find the adequate number of servers for reducing waiting time in an emergency department. They also considered limited waiting room, patients' priorities and single visit of the system by each server within a certain period of time in the model. Agnihothri and Taylor (1991) utilized a $M/M/C$ queueing model to find the optimal staffing levels to handle the variation in call arrivals to an appointment system. They found that the existing staff and the number of hours they were working was enough to handle the demand and by redistributing server capacities over time, they could effectively reduce customer complaints. Green et al. (2006) used queueing analysis to identify provider-staffing patterns in order to minimize LWTR. They showed that despite an increase of 6.3% in patient arrivals, through a weekly 3.1% increase in staffing, LWTR could be decreased by 22.9%. Cochran et al. (2009) used queueing models to find the staffing levels of different emergency department areas.

Patient Planning

Patient appointment systems are highly correlated with patients' waiting time and server utilization. Since decreasing patients' waiting time and increasing expensive servers' utilization are primary goals of a healthcare provider, it is necessary for them to design and implement an efficient appointment system to be able to improve the quality of their service. According to the study done by Huang (1994), the major reason for patients' complaints about their experiences of visiting a healthcare facility is long waiting times. Baily (1952) studied the appointment system and queueing process of a hospital outpatient department. In this work, the author suggested that the patients' appointments should be given at a regular interval, each equal to average service time and the server starts working when the second patient arrives. The impact of variation in demand, size of the queue and appointment intervals was also studied. Baily (1954) used queueing models to find the number of beds and servers in a hospital. Considering a defined amount of waiting, he also determined an appointment system to assign patients to servers in this hospital. Mercer (1960) considered a more general single server appointment system in which customers are scheduled in identical time intervals but may arrive late. The lateness in this work was modeled using a general distribution. Jansson (1966) studied single server systems with constant inter-arrival times and exponential service times ($D/M/1$). The cost of the system was defined as the linear sum of the customers waiting time and the time that the server is idle. After evaluating the total system cost distribution for the k^{th} customer, the author showed that it could be minimized by assigning proper constant inter-arrival time and the initial number of customers in the system. Mercer (1973) extended his previous work by considering bulk arrivals and single arrivals with general service time distribution. It was concluded that in a queueing system with customers being scheduled, the distribution of the queue length in steady state is a truncated sum of weighted geometric distributions.

Wang (1993) studied static and dynamic scheduling problems in a single-server system in which customers arrive with appointment. The author used a set of nonlinear equations to minimize the weighted customer delay and the server completion time. Static scheduling problem is to schedule a finite number of customer arrivals while there is no scheduled customer in the system. On the other hand, dynamic scheduling problem deals with scheduling only one customer arrival assuming that the system already has a number of scheduled customers. Later Wang (1997) extended the previous work to cover systems in which service times are independent and identically distributed with a Coxian-type distribution. The author derived recursive expressions for the customer flow-time distribution by using phase-type distribution functions and matrix algebraic manipulation. Optimal customer appointments were obtained by integrating this procedure with a nonlinear program. Lau and Lau (2000) addressed the problem of minimizing total system cost in scheduling outpatient appointments and developed a fast and accurate procedure to compute this cost for any given appointment schedule. It was assumed that the service time distributions can be any general distribution, patients are never late and patients waiting times were not considered in total system cost evaluation. They used the proposed procedure to find the optimal appointment schedule with the lowest cost for any given job sequence. DeLaurentis et al. (2006) used queueing networks and simulation to study an open access appointment scheduling system at an urban outpatient clinic. They considered a same-day appointment system and showed that the pre-scheduling horizon and the percentage of patients using open access scheduling are key factors for a successful open access scheduling policy. They also suggested that clinics with many visiting doctors, such as residents, are not good candidates for the same-day open access appointment system. Min and Yih (2010) studied the problem of scheduling patients with different priorities in a healthcare facility with finite capacity. They formulated this problem by using a stochastic dynamic programming model and introduced a structural analysis to find the bounds on the feasible actions. These bounds were used to develop a computationally efficient algorithm for solving this problem. Finally, the patients' priority levels are determined based on the trade-offs between computation time and solution quality.

CONCLUSIONS

In this paper, applications of queueing theory in modeling hospital processes have been reviewed and categorized. Since healthcare facilities are directly dealing with human lives, improving system performance is a very important goal. Increasing servers' utilization and decreasing patients' waiting time can enhance system productivity. Queueing theory provides an effective and powerful modeling technique that can help managers achieve the aforementioned goals. This approach can be easily implemented and has several advantages such as providing good and rapid estimations of the system performance. According to this survey, there are several future research opportunities.

- Studying the effect of bed and nurse flexibility. Beds can be used with different categories of patients and nursing staff can be cross-trained to gain the flexibility to serve different patient categories.
- Developing efficient dynamic priority rules for expensive shared healthcare facilities.
- Developing dynamic resource allocation models in case of multiple patient categories and several service resources.
- Regional capacity allocation based on the characteristics and criticality of the candidate regions.
- Staff and patient planning models are designed for a specific environment and there is no general framework that works well in any environment. Developing a scheme for selecting the best architecture for the appointment system for a given healthcare facility could be a potential future research opportunity in this area.

REFERENCES

1. Agnihothri, S.R. and Taylor P.F. (1991) Staffing a centralized appointment scheduling department in Lourdes Hospital, *Interfaces*, 21, 5, 1–11.
2. Bailey, N.T.J. (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting times, *Journal of the Royal Statistical Society* , 14, 2, :185–199.
3. Bailey, N.T.J. (1954) Queueing for medical care, *Journal of the Royal Statistical Society*, 3, 3, 137–145.
4. Blair, E.L. and Lawrence, C.E. (1981) A queueing network approach to health care planning with an application to burn care in New York state, *Socio-economic Planning Sciences*, 15, 5, 207–216.
5. Broyles, J.R. and Cochran, J.K. (2007) Estimating business loss to a hospital emergency department from patient reneging by queueing-based regression, in *Proceedings of the 2007 IERC*, 613–618.
6. Bruin, A.M., Koole, G.M. and Visser, M.C. (2005) Bottleneck analysis of emergency cardiac in-patient flow in a university setting: an application of queueing theory, *Clinical and Investigative Medicine*, 28, 6, 316–317.
7. Burwell, T.H., Jarvis, J.P. and Mcknew, M.A. (1993) Modeling co-located servers and dispatch ties in the hypercube model, *Computers and Operations Research*, 20, 2, 113–119.
8. Burwell, T.H., Mcknew, M.A. and Jarvis, J.P. (1992) An application of a spatially distributed queueing model to an ambulance system, *Socio-Economic Planning Sciences*, 26, 4, 289-300.
9. Cochran, J. K. and Bharti, A. (2006) A multi-stage stochastic methodology for whole hospital bed planning under peak loading, *International Journal of Industrial and Systems Engineering*, 1, 1/2, 8–36.
10. Cochran, J. K. and Roche, K. T. (2009) A multi-class queueing network analysis methodology for improving hospital emergency department performance, *Computers & Operations Research*, 36, 1497–1512.
11. DeLaurentis, P.C., Kopach, R., Rardin, R., Lawley, M., Muthuraman, K., Wan, H., Ozsen, L. and Intrevado, P. (2006) Open access appointment scheduling - an experience at a community clinic, *IIE Annual Conference and Exposition*.
12. Fiems, D., Koole, G. and Nain, P. (2007) Waiting times of scheduled patients in the presence of emergency requests, *working paper*, Available online at: http://www.math.vu.nl/ koole/articles/report05a/art.pdf.
13. Green, L.V., Soares, J., Giglio, J.F. and Green, R.A. (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing, *Annals of Emergency Medicine*, 13, 1, 61–68.
14. Gross, D. and Harris, C. M. (1998) Fundamentals of Queueing Theory, Wiley-Interscience, 3rd edition, USA.
15. Haussmann, R.K.D. (1970) Waiting time as an index of quality of nursing care, *Health Services Research*, 5, 2, 92–105.

16. Huang, X.M. (1994) Patient attitude towards waiting in an outpatient clinic and its applications, *Health Services Management Research*, 7, 1, 2–8.
17. Iannoni, A.P. and Morabito, R. (2007) A multiple dispatch and partial backup hypercube Queueing model to analyze emergency medical systems on highways, *Transportation Research*, Part E, 43, 755–771.
18. Jansson, B. (1966) Choosing a good appointment system: A study of queues of the type (D/M/1), *OR*, 14, 2, 292–312.
19. Kao, E. P. C. and Tung, G. G. (1981) Bed allocation in a public health care delivery system, *Mgmt. Sci.*, 27, 5, 507–520.
20. Koizumi, N., Kuno, E. and Smith, T.E. (2005) Modeling patient flows using a queueing network with blocking, *Health Care Management Science*, 8, 49–60.
21. Larson, R.C. (1974) A hypercube queueing model for facility location and redistricting in urban emergency services, *Computers and Operations Research*, 1, 67–95.
22. Larson, R.C. (1975) Approximating the performance of urban emergency service systems, *OR*, 23, 5, 845–868.
23. Lau, H.S. and Lau A.H.L. (2000) A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities, *IIE Transactions*, 32, 833–839.
24. McManus, M.L., Long, M.C., Cooper, A. and Litvak, E. (2004) Queueing theory accurately models the need for critical care resources, *Anesthesiology*, 100, 1271–1276.
25. Mendonca, F.C. and Morabito, R. (2001) Analyzing emergency medical service ambulance deployment on a Brazilian highway using the hypercube model, *Journal of the Operation Research Society*, 52, 261–270.
26. Mercer, A. (1960) A queueing problem in which the arrival times of the customers are scheduled, *Journal of the Royal Statistical Society*, Series B, 22, 1, 108–113.
27. Mercer, A. (1973) Queues with scheduled arrivals: A correction, simplification and extension, *Journal of the Royal Statistical Society*, Series B, 35, 1, 104–116.
28. Min, D. and Yih, Y. (2010) An elective surgery scheduling problem considering patient priority, *Computers & Operations Research*, 37, 1091–1099.
29. Panayiotopoulos, J.C. and Vassilacopoulos G. (1984) Simulating hospital emergency departments queueing systems, *European Journal of Operational Research*, 18, 250–258.
30. Preater, J. (2001) A bibliography of queues in health and medicine, *Keele Mathematics Research Report*.
31. Roche, K.T. and Cochran, J.K. (2007) Improving patient safety by maximizing fast-track benefits in the emergency department a queueing network approach, in *Proc. of the 2007 Industrial Engineering Research Conference*, 619–624.
32. Rosenquist, C.J. (1987) Queueing analysis: A useful planning and management technique for radiology, *Journal of Medical Systems*, 11, 6, 413-419.
33. Shmueli, A., Sprung, C.L. and Kaplan E.H. (2003) Optimizing admissions to an intensive care unit, *Health Care Management Science*, 6, 3, 131–136.
34. Siddharthan, K., Jones, W.J. and Johnson, J.A. (1996) A priority queueing model to reduce waiting times in emergency care, *International Journal of Health Care Quality Assurance*, 9, 5, 10–16.
35. Singer, M. and Donoso, P. (2008) Assessing an ambulance service with queueing theory, *Comp. & OR*, 35, 2549–2560.
36. Singer, M., Donoso, P. and Jara, S. (2002) Fleet configuration subject to stochastic demand: an application in the distribution of liquefied petroleum gas, *Journal of the Operations Research Society*, 53, 961–971.
37. Solberg, L., Asplin, B., Weinick R. and Magid, D. (2003) Emergency department crowding: consensus development of potential measures, *Annals of Emergency Medicine*, 42, 6, 824–834.
38. Viswanadham, N. and Narahari, Y. (2001) Queueing network modeling and lead-time compression of pharmaceutical drug development, *Int. J. Prod. Res.*, 39, 2, 395–412.
39. Wang, P.P. (1993) Static and dynamic scheduling of customer arrivals to a single-server system, *Naval Research Logistics*, 40, 3, 345–360.
40. Wang, P. P. (1997) Optimally scheduling N customer arrival times for a single-server system, *Comp. & OR*, 24,703–716.
41. Worthington, D.J. (1987) Queueing models for hospital waiting lists, *J. of the Opr. Res. Society*, 38, 5, 413–422.
42. Worthington, D.J. (1991) Hospital waiting list management models, *J. of the Opr. Res. Society*, 42, 833–843.
43. Zenios, S. A. (1999) Modeling the transplant waiting list: A queueing model with reneging, *Queueing Sys.*, 31, 239-251.