THE INSTITUTE FOR SYSTEMS RESEARCH

ISR TECHNICAL REPORT 2007-24

A Survey of Queuing Theory Applications in Healthcare

Samuel Fomundam, Jeffrey Herrmann

A SURVEY OF QUEUING THEORY APPLICATIONS IN HEALTHCARE

Samuel F. Fomundam
Department of Mechanical Engineering
University of Maryland, College Park, MD 20742

Jeffrey W. Herrmann
Department of Mechanical Engineering and Institute for Systems Research
University of Maryland, College Park, MD 20742

Abstract

This paper surveys the contributions and applications of queuing theory in the field of healthcare. The paper summarizes a range of queuing theory results in the following areas: waiting time and utilization analysis, system design, and appointment systems. The paper also considers results for systems at different scales, including individual departments (or units), healthcare facilities, and regional healthcare systems. The goal is to provide sufficient information to analysts who are interested in using queuing theory to model a healthcare process and want to locate the details of relevant models.

Introduction

The organizations that care for persons who are ill and injured vary widely in scope and scale, from specialized outpatient clinics to large, urban hospitals to regional healthcare systems. Despite these differences, one can view the healthcare processes that these organizations provide as queuing systems in which patients arrive, wait for service, obtain service, and then depart. The healthcare processes also vary in complexity and scope, but they all consist of a set of activities and procedures (both medical and non-medical) that the patient must undergo in order to receive the needed treatment. The resources (or servers) in these queuing systems are the trained personnel and specialized equipment that these activities and procedures require.

A considerable body of research has shown that queuing theory can be useful in real-world healthcare situations, and some reviews of this work have appeared. McClain (1976) reviews research on models for evaluating the impact of bed assignment policies on utilization, waiting time, and the probability of turning away patients. Nosek and Wilson (2001) review the use of queuing theory in pharmacy applications with particular attention to improving customer satisfaction. Customer satisfaction is improved by predicting and reducing waiting times and adjusting staffing. Preater (2002) presents a brief history of the use of queuing theory in healthcare and points to an extensive bibliography of the research that lists many papers (however, it provides no description of the applications or results). Green (2006a) presents the theory of queuing as applied in healthcare. She discusses the relationship amongst delays, utilization and the number of servers; the basic M/M/s model, its assumptions and extensions; and the applications of the theory to determine the required number of servers.

This paper surveys the contributions and applications of queuing theory in the field of healthcare. The reviews mentioned above focus on presenting mathematical models or limit their scope to a single type of application. This paper, however, seeks to show the applicability of queuing theory from the perspective of healthcare organizations. Thus, this paper summarizes a range of queuing theory results in the following areas: waiting time and utilization analysis, system design, and appointment systems. This covers processes that provide direct patient treatment and processes that provide auxiliary services such as pharmacy and medical laboratory processing. The paper also considers results for systems at different scales, including individual departments (or units), healthcare facilities, and regional health systems. The goal is to provide sufficient information to analysts who are interested in using queuing theory to model a

healthcare process and want to locate the details of relevant models. We assume that the reader is familiar with healthcare organizations and the basic concepts of queuing theory.

This survey covers analytical queuing theory models applied directly to healthcare systems. It is reasonable for an analyst to understand, adapt, and apply such a model to his own situation. Because they require specialized software and the details of the simulation model are usually unknown, this paper does not review simulation studies of healthcare processes.

Queuing models and simulation models each have their advantages. It is clear that queuing models are simpler, require less data, and provide more generic results than simulation (see also Green, 2006a). However, discrete-event simulation permits modeling the details of complex patient flows. Jacobson et al. (2006) present a list of steps that must be done carefully to model each healthcare scenario successfully using simulation and warn about the slim margins of tolerable error and the effects of such errors in lost lives. Tucker et al. (1999) and Kao and Tung (1981) use simulation to validate, refine or otherwise complement the results obtained by queuing theory. Albin et al. (1990) show how one can use queuing theory for get approximate results and then use simulation models to refine them. We will not explore simulation studies further in this paper.

Spreadsheets and software tools based on queuing theory research can automate the necessary calculations. For example, Albin et al. (1990) use the QNA software, which calculates the time that patients are in a multi-node network, server utilization, the mean and variance of the number of customers at each node, the mean and variance of waiting time at each node, the mean and variance of the number of customers in the network, and the proportion of customers at each node that arrived from other nodes. Aaby et al. (2006) describe the use of spreadsheets to implement queueing network models of mass vaccination and dispensing clinics. However, the

authors are not aware of any other software that is specifically designed to analyze queuing models of healthcare processes.

The next section ("Waiting time and Utilization Analysis") is an overview of research into using queuing theory as an analytical tool to predict how particular healthcare configurations affect delay in patient service and healthcare resource utilization. The "System Design" section reviews research of a prescriptive nature that seeks to determine the optimal allocation of resources necessary to attain the goals determined by healthcare providers and decision makers. In "Appointment Systems" we look at applications to appointment scheduling where the main challenge is reducing patient waiting without greatly increasing server idleness. The "System Size" section considers results for systems of different scales. The paper ends with some observations about the use of queuing theory and suggestions for future research.

Waiting Time and Utilization Analysis

In a queuing system, minimizing the time that customers (in healthcare, patients) have to wait and maximizing the utilization of the servers or resources (in healthcare, doctors, nurses, hospital beds, e.g.) are conflicting goals.

Reneging

When a patient is waiting in a queue, he may decide to forgo the service because he does not wish to wait any longer. This phenomenon, called *reneging*, is an important characteristic of many healthcare systems. The probability that a patient reneges usually increases with the queue length and the patient's estimate of how long he must wait to be served. In systems where demand exceeds server capacity, reneging is the only way that a system attains a "state of dysfunctional equilibrium" (Hall et al., 2006).

An important example of such a system is an emergency department. Broyles and Cochran (2007) calculate the percentage of patients who leave an emergency department without getting help using arrival rate, service rate, utilization, capacity. From this percentage, they determine the resulting revenue loss.

It is possible to redesign a queuing system to reduce reneging. A common approach is to separate patients by the type of service required. Roche et al. (2007) find that the number of patients who leave an emergency department without being served is reduced by separating non-acute patients and treating them in dedicated fast-track areas. Most of their waiting would be for tests or test results after having first seen a doctor. The paper also estimates the size of the waiting area for patients and those accompanying them.

Variable Arrival Rate

Although most analytical queuing models assume a constant customer arrival rate, many healthcare systems have a variable arrival rate. In some cases, the arrival rate may depend upon time but be independent of the system state. For instance, arrival rates change due to the time of day, the day of the week, or the season of the year. In other cases, the arrival rate depends upon the state of the system.

A system with congestion discourages arrivals. Worthington (1991) suggests that increasing service capacity (the traditional method of attempting to reduce long queues) has little effect on queue length because as soon as patients realize that waiting times would reduce, the arrival rate increases, which increases the queue again. Worthington (1987) presents an $M(\lambda q)/G/S$ model for service times of any fixed probability distribution and for arrival rates that decrease linearly with the queue length and the expected waiting time.

The arrival rate may increase over time due to population growth or other factors. Rosenquist (1987) studies how an increase in patient arrival rate affects waiting times and queue length for an emergency radiology service.

Priority Queuing Discipline

In most healthcare settings, unless an appointment system is in place, the queue discipline is either first-in-first-out or a set of patient classes that have different priorities (as in an emergency department, which treats patients with life-threatening injuries before others).

McQuarrie (1983) shows that it is possible, when utilization is high, to minimize waiting times by giving priority to clients who require shorter service times. This rule is a form of the shortest processing time rule that is known to minimize waiting times. It is found infrequently in practice due to the perceived unfairness (unless that class of customers is given a dedicated server, as in supermarket check-out systems) and the difficulty of estimating service times accurately.

When arriving patients are placed in different queues, each of which has a different service priority, the queue discipline may be preemptive or non-preemptive. In the latter, low priority patients receive service only when no high priority patients are waiting, but the low priority patient who is receiving service is not interrupted if a high priority patient arrives and all servers are busy. In the preemptive queue discipline, however, the service to a low priority patient is interrupted in this event. Green (2006a) presents models for both queue disciplines.

Siddhartan et al. (1996) analyze the effect on patient waiting times when primary care patients use the Emergency Department. They propose a priority discipline for different categories of patients and then a first-in-first-out discipline for each category. They find that the

priority discipline reduces the average wait time for all patients: however, while the wait time for higher priority patients reduces, lower priority patients endure a longer average waiting time.

Haussmann (1970) investigates the relationship between the composition of prioritized queues and the number of nurses responding to inpatient demands. The research finds that slight increases in the number of patients assigned to a nurse and/or a patient mix with more high-priority demands result in very large waiting times for low priority patients.

Worthington (1991) analyzes patient transfer from outpatient physicians to inpatient physicians. The patient is assigned one of three priority levels. Based on the priority level, there is a standard time period before which a referred patient should be scheduled to see the inpatient physician. The model assumes sufficient in-patient capacity to treat the highest priority category within its standard time, and proposes sharing the remaining service capacity amongst the lower priority levels in such a manner that they each exceed their standard target times by the same percentage.

Taylor et al. (1969) model an emergency anesthetic department operating with priority queuing discipline. They are interested in the probability that a patient would have to wait more than a certain amount of time to be served.

Fiems et al. (2007) investigate the effect of emergency requests on the waiting times of scheduled patients with deterministic processing times. It is a preemptive repeat priority queuing system in which the emergency patients interrupt the scheduled patients and the latter's service is restarted as opposed to being resumed. This paper models a single server queue and divides time into equally long slots (discretizing time). Periods of emergency interruptions are considered to have no server available from the point of view of the scheduled patients (vacation). The result is a discrete-time queuing model with exhaustive vacations.

Blocking

Blocking occurs when a queuing system places a limit on queue length. For example, an outpatient clinic may turn away walk-in patients when its waiting room is full. In a hospital, where in-patients can wait only in a bed, the limited number of beds may prevent a unit from accepting patients. McManus et al. (2004) present a medical-surgical Intensive Care Unit where critically ill patients cannot be put in a queue and must be turned away when the facility is fully occupied. This is a special case where the queue length cannot be greater than zero, which is called a pure loss model (see Green, 2006a, for more details).

Koizumi et al. (2005) find that blocking in a chain of extended care, residential and assisted housing facilities results in upstream facilities holding patients longer than necessary. They analyze the effect of the capacity in downstream facilities on the queue lengths and waiting times of patients waiting to enter upstream facilities. System-wide congestion could be caused by bottlenecks at only one downstream facility.

System Design

Because patient waiting is undesirable, limiting waiting times is an important objective when designing a healthcare system. This section reviews work on determining system capacity based on desired system goals and requirements. The variables of interest are usually staffing levels, beds, or other key resources.

Bailey (1954) first establishes the existence in outpatient and inpatient clinics of a threshold capacity which occurs at the point where service supply equals demand. When the number of servers is below this threshold, a clinic develops an infinite queue. Slightly above this threshold, waiting time and queue length are low. He argues that it is therefore sufficient to

design for a capacity that exceeds the expected demand (with stochastic error accounted for) by a value of 1 or 2. Long waiting lists are most likely the result of accumulated backlog which can be depleted by a temporary surge in supply. Seasonal variations in supply would also result in a sharp rise in waiting list length.

Moore (1977) reduces customer waiting time for birth and death certificates at the Dallas bureau of vital statistics by decreasing the time required to serve each customer. This research first uses queuing theory to calculate the service rate required to achieve a target waiting time of 15 minutes. This service rate is converted to the time required to serve one customer. The reduced time required to serve each customer is attained through the use of new equipment and more efficient processes.

Agnihothri and Taylor (1991) seek the optimal staffing at a hospital scheduling department that handles phone calls whose intensity varies throughout the day. There are known peak and non-peak periods of the day. The paper groups periods that receive similar call intensity and determines the necessary staffing for each such intensity, so that staffing varies dynamically with call intensity. As a result of redistributing server capacity over time, customer complaints immediately reduced without an addition of staff. Green (2006b) uses the same approach and names it Stationary Independent Period by Period (SIPP) to adjust staffing in order to reduce the percentage of patients that renege. However, arguing that congestion starts some time after the arrival peak, the staffing levels should lag behind the service demand levels (lag SIPP).

Blocking

In systems with blocking, congestion not only increases patient waiting time but also reduces the throughput of the system. Bruin et al. (2005) determine the number of beds required

to achieve a maximum turn away rate of 5% at the emergency cardiac department of the university medical centre of Amsterdam which implements the pure loss model. Cooper and Corcoran (1974) deal with the same problem extended to a sequence of two stations each of which should have a maximum turn-away rate of 5%. Milliken et al. (1972) seek a 1% turn-away rate in an obstetrics department in which vaginal births have priority over scheduled caesarian sections. They point out the benefits of economies of scale so that larger facilities incur lower bed investment per additional birth.

Given a desired maximum turn-away rate, Bruin et al. (2007) determine the optimal number of beds in a cardiology department. The cardiology department is modeled as a network of 3 sub-departments. The research finds that too few beds downstream is the primary cause of refused admissions upstream and that congestion effects can add 20-30% to patient length of stay in the department. They characterize having a fixed target utilization rate as unrealistic and conclude that a downstream utilization of 55% is necessary to attain a 2% turn-away rate. As an alternative, departments could be merged to gain the benefits of economies of scale thereby meeting the goal at higher occupancy rates.

Blair and Lawrence (1981) seek to design the capacity of horizontally integrated burn care facilities throughout the state of New York, so that no more than 5% of patients are turned away from the system. If a patient goes to a facility which is fully occupied, that facility would refer to the patient to another which is not filled. If all facilities are fully occupied, the patient is lost to the system. First, they use queuing theory to determine the capacity of the entire system as if it were one queuing system. This capacity is then allocated to facilities in a manner that best attains their individual goals. They find such a system-planned approach ideal for a system with low demand and high infrastructural costs.

Tucker et al. (1999) consider activating a second operating room (OR) team during the night shift. Using queuing theory, they find that the probability of two patients needing the OR services is negligible.

Minimize Costs

Determining server capacity by minimizing the costs in a healthcare queuing system is a special case of system design. Most of the research assigns costs to patient waiting time and to each server. After modeling the system using queuing theory, minimizing costs reduces to an exercise of finding the resource allocation that costs the least or generates the most profit.

Keller and Laughhunn (1973) set out to determine the capacity with minimal costs required to serve patients at the Duke University Medical center. They find that the current capacity is good but needs to be redistributed in time to accommodate patient arrival patterns.

Young (1962a, b) proposes an incremental analysis approach in which the cost of an additional bed is compared with the benefits it generates. Beds are added until the increased cost equals the benefits.

Shimshak et al. (1981) consider a pharmacy queuing system with preemptive service priority discipline where the arrival of a prescription order suspends the processing of lower priority prescriptions. Different costs are assigned to wait-times for prescriptions of different priorities.

Gupta et al. (1971) choose the number of messengers required to transport patients or specimens in a hospital by assigning costs to the messenger and to the time during which a request is in queue. In this problem, non-routine requests are superimposed on top of routine, scheduled requests. The authors also calculate the number of servers required so that a given

percentage of requests does not exceed a given wait time and the average number of patients in the queue does not exceed a given threshold.

Assuming a phase-type service distribution, Gorunescu et al. (2002a) assign costs based on a base stock inventory policy. In this pure loss model, there is a holding cost associated with an empty bed, a penalty cost associated with each patient turned away, and a profit assigned to each day a bed is occupied.

Khan and Callahan (1993) incorporate advertising into their model to control the demand for laboratory services. For each staffing level, they determine the number of clients that would maximize profits. They then choose the staffing level with maximum profits and apply the necessary amount of advertising that would attract the desired number of clients. The model assumes that clients would leave without service if they wait above a certain amount of time.

Rosenquist (1987) chooses staffing capacity in an outpatient radiology service with a limited waiting area by minimizing cost. He suggests scheduling patients when possible and segregating patients based on expected examination duration. Such measures would reduce variability and decrease expected waiting times.

Gorunescu et al. (2002b) use backup beds (only staffed during peak demand) to reduce the probability of patient turn-away at a marginal cost. The model assumes a phase-type service distribution.

Appointment Systems

Compared to systems without appointments, systems with appointments reduce the arrival variability and waiting times at the facility. However, it is important to note that systems with appointments require patients to wait outside the facility. Of course, because it is not at the facility, this waiting can be productive time and therefore has lower cost to the patient. (Plus,

they do not occupy space in the facility's waiting rooms.) A key issue has been to reduce patient waiting times without causing a significant increase in doctor idle time, a significant cost for the healthcare facility.

Bailey (1952, 1954) proposes (a) appointment interval and (b) consultant arrival time as two variables that determine the efficiency of an appointment system. In order to find a balance between patient wait time and consultant idle time, first determine the relative values of patient time and consultant time. The ratio of the total time wasted by all patients to the consultant's idle time should equal the value of the consultant's time relative to the patients'. He chooses to assign individual appointment times at intervals equal to the average patient processing time and finds that the consultant should arrive at the same time as the second patient.

Brahimi and Worthington (1991) design an appointment system to reduce the number of patients in the queue at any time, and reduce patient waiting time without significantly increasing doctor idle time. They explore the effect of patients who do not show up for their appointments. The clinic starts out with a certain number of patients waiting and a maximum number of patients allowed at any time.

In Vasanawala and Desser (2005) a radiology department has some time slots scheduled for routine radiology analysis. Emergency requests may require rescheduling of scheduled requests. Given a 1% or 5% probability of rescheduling, the authors use queuing theory to determine how many scheduled slots to leave empty during routine scheduling.

Many outpatient appointments allow booking appointments months in advance. DeLaurentis et al. (2006) point out that patient no-shows without cancelling appointments could lead to waste of resources. They propose implementing short-notice appointment systems based on a queuing network analysis tailored to the realities of any particular outpatient clinic. Their

approach assumes the availability of a certain number of staff who can be distributed amongst the different stations of the queuing network in several combinations. A combination is chosen based on its resulting utilization per station and expected patient length of stay in clinic. The implementations of these ideas did not improve the appointment system, a failure which they attribute to the clinic using many visiting doctors and the patients being unable to schedule visits with their primary care physician at short notice.

Bottlenecks

In a queuing network, there are several nodes at which services are dispensed. A patient may have to go through several nodes, and thus several queues in order to obtain the desired service. In the context of appointment systems, we can expect nodes where the ratio of demand to available service capacity is relatively high to become bottlenecks. Such bottlenecks would have high utilization and increase overall patient waiting times even though other nodes may have low utilization.

Albin et al. (1990) find the bottlenecks at the Hurtado Health Center appointment clinic by collecting data and analyzing it using QNA, the queuing network analysis software program. Though their model deviates appreciably from assumptions, they are able to find the bottle necks by identifying the nodes where wait times are longest. They then reduce overall waiting time by offering common-sense recommendations on a node-by-node basis.

System Size

As mentioned in the introduction, the size of healthcare organizations varies greatly. Following Hall et al. (2006), we can distinguish between three different scales. The smallest scale is the department, “a unit within a larger center oriented toward performing a single

function, or a group of closely related functions." The next larger scale organization is the health care center, which is a group of proximate, coordinated departments amongst which patients can flow. The largest scale that we consider is the regional health system, a hierarchy of facilities with the most routine services provided by local clinics and the most specialized, resource-intensive services provided at a few regional facilities.

Hall et al. (2006) also present a macro system scale that considers the life cycle of an individual's state of wellness and his interactions with the healthcare system throughout a lifetime. We have found no research applying queuing theory to this type of system.

Most of the research reviewed above has been done at the department scale. Here we will highlight some work at the two larger scales.

Kao and Tung (1981) investigate the redistribution of hospital beds amongst the inpatient departments of a hospital. First, a baseline patient capacity is chosen for each department. Additional beds are then allocated to departments in a manner to minimize patient overflows from one department to another. Forecasts are used to determine both the baseline bed allocation and the anticipated patient demand in order to minimize overflow.

Blair and Lawrence (1981) investigate a regional hierarchy of burn care facilities where excess demand at one facility is absorbed by other facilities in the same region and overflows at one region are absorbed by other regions. Worthington (1991) considers the coordination of patient flow from outpatient clinics to inpatient clinics. Koizumi et al. (2005) model the mental healthcare system as a chain of facilities including acute hospitals, extended acute hospitals, residential facilities and supported housing.

Conclusions

This paper has surveyed the use of queuing theory for the analysis of different types of healthcare processes. Models for estimating waiting time and utilization, models for system design, and models for evaluating appointment systems have been presented. The survey has reviewed models for departments (or units), facilities, and systems.

We can draw some conclusions from the work surveyed above. The variability in demand for healthcare services and service times mean that simplistic rules like mandating specific utilization levels or fixing patient to resource ratios would lead only to congestion and poor quality of service and are unlikely to be successful approaches to contain or reduce healthcare costs. Larger organizations with more patients are able to attain the same quality of service at higher utilizations than smaller organizations. Although appointment systems are often designed to avoid doctor idle time (without considering patient waiting time), it is possible to reduce patient wait time without significantly increasing doctor idle time.

As long as increasing the productivity of healthcare organizations remains important, analysts will seek to apply relevant models to improve the performance of healthcare processes. This paper shows that many models are available today. However, analysts will increasingly need to consider the ways in which distinct queuing systems within an organization interact. Consider, for instance, the following analogy from manufacturing, where a traditional factory (with a functional layout) has been transformed into manufacturing cells whose production is closely linked to the final assembly line through simple signals (such as kanbans). While healthcare organizations don't resemble factories, they do have links between subsystems (such as operating rooms and the post-operative recovery unit), and these interfaces need

improvements. Developing appropriate models of the links (or interfaces) between the distinct queuing systems is an important direction for future research.

Acknowledgements

Cooperative Agreement Number U50/CCU302718 from the CDC to NACCHO supported this publication. Its contents are solely the responsibility of the University of Maryland and the Advanced Practice Center for Public Health Emergency Preparedness and Response of Montgomery County, Maryland, and do not necessarily represent the official views of CDC or NACCHO.

References

Aaby, K., Herrmann, J.W., Jordan, C., Treadwell, M., and Wood, K. (2006) Using operations research to improve mass dispensing and vaccination clinic planning. Interfaces, 36, 569-579.

Agnihothri, S.R. and Taylor P.F. (1991) Staffing a centralized appointment scheduling department in Lourdes Hospital. Interfaces 21, 1-11.

Albin, S.L., Barrett, J., Ito, D. and Mueller, J.E. (1990) A queueing network analysis of a health center. Queueing Systems, 7, 51-61.

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, 185-199.

Bailey, N.T.J. (1954) Queuing for medical care. Applied Statistics, 3, 137-145.

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, 207-216.

Brahimi, M. and Worthington, D.J. (1991) Queueing models for out-patient appointment systems – a case study. The Journal of the Operational Research Society, 42, 733-746.

Broyles, J.R. and Cochran, J.K. (2007) Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression, in Proceedings of the 2007 Industial Engineering Research Conference, 613-618.

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, 316-317.

Bruin, A.M., Rossum, A.C., Visser, M.C. and Koole, G.M. (2007) Modeling the emergency cardiac in-patient flow: an application of queuing theory. Health Care Management Science, 10, 125-137.

Cooper, J.K. and Corcoran, T.M. (1974) Estimating bed needs by means of queuing theory. The New England Journal of Medicine, 291, 404-405.

DeLaurentis, P., 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, in IIE Annual Conference and Exposition.

Fiems, D., Koole, G. and Nain, P. (2007) Waiting times of scheduled patients in the presence of emergency requests. http://www.math.vu.nl/~koole/articles/report05a/art.pdf, accessed August 6, 2007.

Gorunescu, F., McClean, S.I. and Millard, P.H. (2002a) A queueing model for bed-occupancy management and planning hospitals. Journal of the Operational Research Society, 53, 19-24.

Gorunescu, F., McClean, S.I. and Millard, P.H. (2002b) Using a queueing model to help plan bed allocation in a department of geriatric medicine. Health Care Management Science, 5, 307-312.

Green, L. (2006a) Queueing analysis in healthcare, in Patient Flow: Reducing Delay in Healthcare Delivery, Hall, R.W., ed., Springer, New York, 281-308.

Green, L.V. (2006b) Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine, 13, 61-68.

Gupta, I., Zoreda, J. and Kramer, N. (1971) Hospital manpower planning by use of queueing theory. Health Services Research, 6, 76-82.

Hall, R., Belson, D., Murali, P. and Dessouky, M. (2006) Modeling patient flows through the healthcare system, in Patient Flow: Reducing Delay in Healthcare Delivery, Hall, R.W. ed., Springer, New York, 1-44.

Haussmann, R.K.D. (1970) Waiting time as an index quality of nursing care. Health Services Research, 5, 92-105.

Jacobson, S., Hall, S. and Swisher, J. (2006). Discrete-event simulation of health care systems, in Patient Flow: Reducing Delay in Healthcare Delivery, Hall, R.W. ed., Springer, New York, 211-252.

Khan, M.R. and Callahan, B.B. (1993) Planning laboratory staffing with a queueing model. European journal of operational research, 67, 321-331.

Kao, E.P.C. and Tung, G.G. (1981) Bed allocation in a public health care delivery system. Management Science, 27, 507-520.

Keller, T.F. and Laughhunn, D.J. (1973) An application of queuing theory to a congestion problem in an outpatient clinic. Decision Sciences, 4, 379-394.

Koizumi N., Kuno, E.and Smith, T.E. (2005) Modeling patient flows using a queuing network with blocking. Health Care Management Science, 8, 49-60.

McClain, J.O. (1976) Bed planning using queuing theory models of hospital occupancy: a sensitivity analysis. Inquiry, 13, 167-176.

McManus, M.L., Long, M.C., Cooper, A. and Litvak, E. (2004) Queuing theory accurately models the need for critical care resources. Anesthesiology, 100, 1271-1276.

McQuarrie D.G. (1983) Hospital utilization levels. The application of queuing theory to a controversial medical economic problem. Minnesota Medicine, 66, 679-86.

Milliken, R.A., Rosenberg, L. and Milliken, G.M. (1972) A queuing model for the prediction of delivery room utilization. American Journal of Obstetrics and Gynecology, 114, 691-699.

Moore, B.J. (1977) Use of queueing theory for problem solution in Dallas, Tex., Bureau of Vital Statistics. Public Health Reports, 92, 171-175.

Nosek, Jr., R.A. and Wilson, J.P. (2001) Queuing theory and customer satisfaction: a review of terminology, trends, and applications to pharmacy practice. Hospital Pharmacy, 36, 275-279.

Preater, J. (2002) Queues in health. Health Care Management Science, 5, 283.

Roche, K.T., Cochran, J.K. and Fulton, I.A. (2007) Improving patient safety by maximizing fast-track benefits in the emergency department – a queuing network approach, in Proceedings of the 2007 Industrial Engineering Research Conference, pp 619-624.

Rosenquist, C.J. (1987) Queueing analysis: a useful planning and management technique for radiology. Journal of Medical Systems, 11, 413-419.

Shimshak, D.G., Gropp Damico, D. and Burden, H.D. (1981) A priority queuing model of a hospital pharmacy unit. European Journal of Operational Research, 7, 350-354.

Siddhartan, K., Jones, W.J. and Johnson, J.A. (1996) A priority queuing model to reduce waiting times in emergency care. International Journal of Health Care Quality Assurance, 9, 10-16.

Taylor, T.H., Jennings, A.M.C., Nightingale, D.A., Barber, B., Leivers, D., Styles, M. and Magner, J. (1969) A study of anaesthetic emergency work. Paper 1: The method of study and introduction of queuing theory. British Journal of Anaesthesia, 41, 70-75.

Tucker, J.B., Barone, J.E., Cecere, J., Blabey, R.G. and Rha, C. (1999) Using queueing theory to determine operating room staffing needs. Journal of Trauma, 46, 71-79.

Vasanawala, S.S. and Desser, T.S. (2005) Accommodation of requests for emergency US and CT: Applications of queueing theory to scheduling of urgent studies. Radiology, 235, 244-249.

Worthington, D.J. (1987) Queueing Models for Hospital Waiting Lists. The Journal of the Operation Research Society 38, 413-422.

Worthington, D. (1991) Hospital waiting list management models. The Journal of the Operational Research Society 42, 833-843.

Young J.P. (1962a) The basic models, in A Queuing theory approach to the control of hospital inpatient census, John Hopkins University, Baltimore, 74-97.

Young J.P. (1962b) Estimating bed requirements, in A Queuing theory approach to the control of hospital inpatient census, John Hopkins University, Baltimore, 98-108.

Biographical Sketches

Samuel Fomundam is a graduate research assistant at the University of Maryland. He earned his B.S. in computer engineering from the United States Military Academy at West Point in 2000.

Jeffrey W. Herrmann is an associate professor at the University of Maryland, where he holds a joint appointment with the Department of Mechanical Engineering and the Institute for Systems Research. He is the director of the Computer Integrated Manufacturing Laboratory and Associate Director for the University of Maryland Quality Enhancement Systems and Teams (QUEST) Honors Fellows Program. He is a member of INFORMS, ASME, IIE, SME, and ASEE.

Dr. Herrmann earned his B.S. in applied mathematics from Georgia Institute of Technology. As a National Science Foundation Graduate Research Fellow from 1990 to 1993, he received his Ph.D. in industrial and systems engineering from the University of Florida. His dissertation investigated production scheduling problems motivated by semiconductor manufacturing.