International Journal of Information Technology and Knowledge Management
Vol. 5 No. 1 (January-June, 2019)

Received: 09th January 2019 *Revised: 18th February 2019* *Accepted: 20 June 2019*

Corresponding Reviewer: Dennis Bams

Estimating theVariance of Waiting Time in the Delivery of Health Care Services

Xiaofeng Zhao* and Kenneth Gilbert**

**College of Business, The University of Mary Washington, Fredericksburg, VA 22401*
***College of Business Administration, The University of Tennessee, Knoxville, TA, 37919*

Authors emails
xzhao@umw.edu
kgilber1@utk.edu

ABSTRACT

Waiting lines and delays have become commonplace in the healthcare industry. As a result, appointment system is widely used to improvements in patient satisfaction. To implement appointment scheduling, we need to know not only the average of customer waiting time, but also the variance around average waiting time. This research provides an approximation method to measure the variance of waiting time in the general queue, which requires only the specification of the mean and the standard deviation of the inter-arrival and service times. It can be easily implemented in a spreadsheet and applied to healthcare operations. The results demonstrate the usefulness of the queuing models in providing guidance on implementing appointment scheduling and waiting time guarantee strategy.

***Keywords**: Queuing theory, Stochastic processes,Healthcare scheduling,Waiting time,ApproximationMethods*

INTRODUCTION

Health care practices are increasingly competing not only on cost, but also on quality and patient satisfaction. As other service industries must do, healthcare organizations must strive to balance customer demands for better service while simultaneously controlling the costs of providing service. In this environment, timely access to care has become a more important issue.As a quality issue, excessive delays in scheduling tests or procedures can result in delayed diagnostic information or deterioration in the patient's condition. As a result, physician practices are eager to embrace new approaches to patient appointment scheduling to reduce backlogs, increase productivity, and improve patient satisfaction. To this end, many healthcare organizations have taken steps to improve quality by adapting appointment scheduling and waiting time guarantee strategy.

In theory, an appointment system reduces patient waiting time. In practice, the waiting time can still be substantial. Outpatient appointment scheduling in health care has been researched over the last 50 years (Bailey 1952, Green and Savin 2008). Various scheduling rules have been proposed in different research works. A good appointment schedule is one that trade-offs patients' waiting time for clinics' overtime, constrained by the patient load and staffing.

One simple guideline from these studies is to place cases with low variability of consult duration in the beginning of the session. Typically, first visit patients have a higher variation in consult duration than follow-up patients do. Hence, the guideline suggests placing the follow-up patients in the beginning of the session. The clinics can further adjust this guideline according to their patients' characteristics.

Unfortunately, such variation may be overlooked or trivialized if the phenomenon is not well understood by healthcare managers. Knowing how variation affects the delivery of services creates opportunities for focused improvement. Currently, these practices have no

guidelines or frameworks to help identify an appropriate balance between physician capacity and patient panel sizes that are consistent with manageable patient backlogs. A critical starting point for any organization striving to improve service is recognition and understanding of variation of waiting time.Both theoretical results and practical case studies have demonstrated how variation in the arrival process and in the delivery of service can cause delays(Hoppand Spearman 2000, Noon *et al.* 2003).

In many systems, the "worst case" value of patient flow time is very relevant because it represents the turnaround time that can safely be promised to the customers. Predicting the range of variation of the time in the system (rather than just the average) is needed for healthcare decision-making. To implement appointment scheduling, we need to know not only the average of customer waiting time, but also the variance around average waiting time. Knowing the variance of the patient waiting time is essential to understanding the performance of queuing system in healthcare delivery.

Research on patient waiting time has traditionally been the domain of queuing theory. The organizations that care for people 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 (Aaby *et al.* 2006, Griffiths *et al.* 2006). 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.

Queues occur because of uncertainty in the environment; whenever the demand for service exceeds the ability to provide service, a queue forms. Essentially, queues arise when service is demanded while a server is busy providing service to others. In healthcare, queues are commonplace at registration desks, walk-in clinics, and emergency rooms. Queues also exist without the telltale line, such as when people are on hold when calling for appointments, referrals, or prescription refills or when patients are waiting for a bed transport or housekeeping services for their rooms. The demand for healthcare services typically originates in a random fashion. Accidents that result in a trip to an emergency room, symptoms that result in a visit to the doctor, and the decision to receive flu shot all lead to a random arrival process. Likewise, the delivery of service is also subject to variation. For example, the time it takes to draw a sample of blood depends on such things as the availability of the appropriate kit, the skill and training of the healthcare worker, and the condition of the patient (Preater 2002, Bennett 1998).

Queues differ according to various characteristics that distinguish them from one another. A major distinction classifies queues according to the number of servers and the distributions that characterize the arrival rates of customers (or their inter-arrival times) and the service times. From a statistical perspective, the random arrival process is not necessary described with the Poisson probability distribution. Similarly, the exponential probability distribution is inappropriate when a wide range of service times is possible (Hopp and Spearman 2000). Kendall notation A/B/n is widely accepted in queuing system. In this notation, the A, B, and n denote, respectively, the inter-arrival time distribution, the service time distribution and the number of servers. In other words, most health care queuing problems are the general GI/G/n system(G for general, I for independent arrivals).

Unfortunately, without the memory-less property of the exponential distribution to facilitate analysis, we cannot compute exact performance measures for the GI/G/n queue. When it comes to exact solutions of multi-server queuing systems, the more one departs from the assumption of exponential, the thornier the problem becomes, especially if this happens for the service time. Due to its inherent complexity, analysis of the GI/G/n queue in general is extremely difficult (Bertsimas 1990, Whitt 1993, 2004).

In this research, we provide approximation methods for the variance (standard deviation) of waiting timefor a general multi-server queue with infinite waiting capacity GI/G/n.The approximations require only the mean and standard deviation or the coefficient of variation of the inter-arrival and service time distributions, and the number of servers. These approximations are simple enough to be implemented in spreadsheet calculations, but in comparisons to Monte Carlo simulations have proven to give good approximations (within ± 10%) for cases in which the coefficients of variation for the inter-arrival and service times are between 0 and 1. 25. The approximations also have the desirable properties of being exact for the specific case of Markov queue modelM/M/n, as well as some imbedded Markov chain queuing models.The spreadsheet can be easily applied to healthcare operations to calculate the variance of waiting time. The results demonstrate the usefulness

of the queuing models in providing guidance on implementing appointment scheduling and waiting time guarantee strategy. Another feature of our model is that managers can conduct what-if analysis and select appropriate capacity levels so as to commit themselves to a given waiting-time guarantee. Hopefully the approximation will be beneficial to practitioners in helping them provide simple, quick, and practical answers to their multi-server queuing systems.

The rest of this paper is organized as follows. In section 2, we review the contributions and applications of queuing theory in the field of healthcare. In section 3, we derive exact expression for the coefficient of variation of waiting time for G/M/n and M/G/1 queues. In section 4, we develop interpolation approximation for variance of waiting time for the general queue GI/G/n. In section 5, numerical results show that the approximations are accurate enough to be applied to service operations. Section 6 delivers concluding remarks.

LITERATURE REVIEW

To help healthcare managers evaluate queuing phenomena, a wide variety of analytic and simulated queuing models are available (Kleinrock 1976, Gross and Harris 2002,). In fact, healthcare continues to be one of the fruitful sources of queuing applications. Some of the earliest work was carried out by Bailey (1952) and Welch (1964) in modeling appointment systems in outpatient facilities. 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. Cox, Birchall, and Wong (1985), and Green and Nguyen (2001) provided applications where the complexity of the healthcare system required additional modeling considerations such as the use of mathematical approximations or simulation. 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. Fomundam (2007) summarizes a range of queuing theory results in the following areas: waiting time and utilization analysis, system design, and appointment systems. 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. Green (2006a, 2008) 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.

The research mentioned above survey the contributions and applications of queuing theory in the field of healthcare, showing the applicability of queuing theory from the perspective of healthcare organizations. Queuing models and simulation models each have their advantages. Discrete-event simulation permits modeling the details of complexity 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. 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.

Spreadsheets and software tools based on queuing theory research can automate the necessary calculations. For instance, Aaby et al. (2006) describe the use of spreadsheets to implement queuing network models of mass vaccination and dispensing clinics. It is clear that queuing models are simpler and practical, require less data, and provide more generic results than simulation (Albin, 1990, Green, 2006a). In this paper, we intend to provide a practical spreadsheet solution to the variation of waiting time. Simulation experiments are conducted to test the approximation results.

A considerable body of research has shown that queuing theory can be useful in real-world healthcare situations. However, from a statistics perspective, GI/G/n model and applications are not discussed in the above literature and authors are not aware of any other spreadsheet model that is specifically designed to analyze GI/G/n queuing model of healthcare processes.

In fact, recent years have witnessed a growing volume of good quality approximations for the GI/G/n queue (Whitt 1999, 2004, Atkinson 2008). While the accuracy of these approximations is usually satisfactory, they often result in algebraically intractable expression. This hinders attempts to derive closed-form solutions to the decision variables incorporated in optimization models, and inevitably leads to the use of complex numerical methods or to recursive schemes of calculation. Furthermore, actual application of many of these approximations is often obstructed due to the thorough specification that is needed of inter-arrival or service time distribution (Shore 1988).

In addition, all current literature focuses on the probability of waiting and the average waiting time. The analysis of the variance of waiting time remains unsolved due to its inherent complexity. There is no

mathematically tractable general formula for approximating the standard deviation of waiting time σ_q in the GI/G/nqueue. Only bounds or approximations of waiting time have been found in the literature. When these bounds are used as approximations, they appear to be rather crude (Bertsimas1990, Witt 2004).

ANALYTICAL MODLE DEVELOPMENT

To develop the approximation of the standard deviation of waiting time,we have studied the equivalent problem of finding a mathematically tractable formula of estimating the coefficient of variation of waiting time $c_q = \sigma_q / W_q$, where W_q and σ_q are respectively the average and standard deviation of the time in queue. There exist good approximations for the average waiting time (Kimura 1986, Whitt 1993). For instance, Sakasegawa (1977) presented the following closed-form expression for the mean waiting time in GI/G/n queue:

$$W_q(GI/G/n) = \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho^{\sqrt{2(n+1)}-1}}{n(1-\rho)}\right)\left(\frac{1}{\mu}\right)$$

c_a isthe coefficient of variation of inter-arrival time and c_s coefficient of variation of service time.

The advantage of this formula is discussed in more details (Whitt 1993). Although it may appear complicated, it does not require any type of iterative algorithm to solve and is therefore easily implemented in a spreadsheet program. This also makes it possible to couple the single-station approximation with the multiple-server to create a spreadsheet tool for analyzing the performance of a series of queues. The above formula is used in our research when calculating average customer waiting time for GI/G/n queue.

We present a general expression for c_q which is applicable to G/M/n and M/G/1 queues. We conjecture that this expression provides a good approximation for GI/G/n queues and have tested this conjecture via computer simulations. In the following, λ is the arrival rate, and μ is the service rate of each server and ρ = λ/(μ*n*)

The expression requires as input the third moment of the service time and $P(T_q = 0)$the probability of no waiting. We show that the expression is relative insensitive to small errors in estimating these two parameters and propose rudimentary approximations to these two parameters.

For G/M/n and M/G/1 queues:

$$c_q = \sqrt{1 + \frac{4E[s^3]}{3\lambda}\frac{P(T_q = 0)}{(E[s^2])^2}} \qquad (1)$$

Where $P(T_q = 0)$ is the probability of no waiting, $E[s^2]$ and $E[s^3]$ are the second and third moments of the service time distribution.

Proof: For M/G/1queue, we know the variance of waiting time is $\sigma_q^2 = W_q^2 + \frac{\lambda E[s^3]}{3(1-\rho)}$ and the average waiting time is $W_q = \frac{\lambda E[s^2]}{2(1-\rho)}$ (Kleinrock 1976), where $E[s^2]$, $E[s^3]$ are the second and the third moments of the service time distribution. For M/G/1, we know $P(T_q = 0) = 1 - P(T_q > 0) = 1 - \rho$. Therefore,

$$c_q = \frac{\sigma_q}{W_q}\sqrt{1 + \frac{\lambda E[s^3]}{3(1-\rho)W_q^2}}\sqrt{1 + \frac{4E[s^3]}{3\lambda}\frac{P(T_q = 0))}{(E[s^2])^2}}$$

When general distribution is exponential, M/G/1reduces to M/M/1. ForM/M/1, we know $E[s^2] = 2/\mu^2$ and $E[s^3] = 6/\mu^3$, so expression (1) can be simplified to:

$$c_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}} \qquad (2)$$

For G/M/n queue, we know the distribution of waiting time is

$W_q(t) = 1 - \frac{Cr^n}{1-r}e^{-n\mu(1-r)t}$ $(t \geq 0)$ (Gross and Harris 2002). So we have,

$$W_q = E[T_q] = \int_0^\infty t dW_q(t)$$
$$= \int_0^\infty t\frac{Cr^n}{1-r}e^{-n\mu(1-r)t}n\mu(1-r)dt$$
$$= \frac{Cr^n}{n\mu(1-r)^2}.$$

$$E[T_q^2] = \int_0^\infty t^2\frac{Cr^n}{1-r}e^{-n\mu(1-r)t}n\mu(1-r)dt$$
$$= \frac{2Cr^n}{n^2\mu^2(1-r)^3}.$$

Hence by definition,

$$\sigma_q^2 = E[T_q^2] - (E[T_q])^2 = \frac{2Cr^n(1-r) - C^2r^{2n}}{n^2\mu^2(1-r)^4}.$$

For G/M/n, we want to verify

$$\sigma_q^2 = \frac{2 - P(T_q > 0)}{P(T_q > 0)} \cdot W_q$$

Equivalently,

$$\frac{\sigma_q^2}{W_q^2} = \frac{2 - P(T_q > 0)}{P(T_q > 0)}.$$

$$\text{LHS} = \left(\frac{2Cr^n(1-r) - C^2 r^{2n}}{n^2\mu^2(1-r)^4} \right) \Bigg/ \left(\frac{Cr^n}{n\mu(1-r)^2} \right)^2$$

$$= \frac{2}{Cr^n/(1-r)} - 1 = \frac{2}{P(T_q > 0)} - 1.$$

$$\text{RHS} = \frac{2 - P(T_q > 0)}{P(T_q > 0)} = \frac{2}{P(T_q > 0)} - 1.$$

Therefore, LHS = RHS, $\frac{\sigma_q^2}{W_q^2} = \frac{2 - P(T_q > 0)}{P(T_q > 0)}$;

Hence, for G/M/n & M/M/1, $c_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}}$

Since M/M/1 is a subset of M/G/1, we conclude G/M/n is a special case of M/G/1, when calculating the coefficient of variation of waiting time.

Thus formula (1) holds for G/M/n queues. The above relationship does not depend at all on the inter-arrival time distribution or the number of servers s. This implies that for G/M/n queues, all of the needed information about the inter-arrival time distribution and the number of servers is contained in the probability of waiting $P(T_q > 0)$.

APPLICATION TO GI/G/N QUEUE

We conjecture that formula (1) can be used as an approximation for the GI/G/n queue since it applies to G/M/nand M/G/1.Whitt (1993) conjectured that the exact formula for the distribution of waiting times of M/G/1 can be used as an approximation for the M/G/n model. Seelan and Tijms (1984) provided additional support for this approximation.

To estimate c_q using formula (1), it is necessary to estimate $P(T_q = 0)$. Since we do not assume that $E[s^3]$ is specified, we must also estimate it by assuming some known distribution for the service times, e.g. Weilbull, uniform, or gamma, for which the third moment can be computed as a function of the average and standard deviation.

We analyze the sensitivity of the estimate of c_q to errors in estimating $E[s^3]$ and $P(T_q = 0)$. Expressed as percentage the error in estimating c_q is always less than half of the error in estimating either of these two parameters. Furthermore, from formula (1), it can be seen as $P(T_q = 0)$ approaches 0, the coefficient c_q approaches 1. This means for queues the more congested the queue, the less the impact of the third moment of the service time.

4.1.Estimation of $E[s^3]$

To implement the approximations in a spreadsheet format, we assumed that the service time distribution could be approximated using a gamma distribution with mean of 1/μ, shape parameter α, and scale parameter β. Weestimate the parameters as $\alpha = 1/c_s^2$ and $\beta = 1/\alpha\mu$ where c_s is the coefficient of variation of the service time distribution.

Then: $E[s^2] = M_s^{''}(t)\big|_{t=0} = \alpha(\alpha+1)\beta^2$.

$E[s^3] = M_s^{'''}(t)\big|_{t=0} = \alpha(\alpha+1)(\alpha+2)\beta^3$

Then substitute these values into formula (1) we have:

$$c_q = \sqrt{1 + \frac{4(1 - P(T_q > 0)(\alpha + 2)}{3P(T_q > 0)(\alpha + 1)}} \qquad (3)$$

4.2.Estimation of $P(T_q > 0)$

Estimating the probability of waiting for GI/G/n is a complicated and tedious problem. Whitt (2004) discussed this problem in more details. His results show that probability of waiting is related to coefficients of variations of inter-arrival times and service times. It depends much more on inter-arrival times. In this research, we develop an interpolation method for estimating the probability of waiting $P(T_q > 0)$ in the GI/G/n queue that gives exact results for M/M/n, E_k/M/1, and M/G/1 queues. The result is consistent with Whitt's conclusion.

For a multi-server queue GI/G/n, we first approximate it via a single server queue. We compute $P(T_q > 0)$ for M/M/n queue having the same arrival rate and service rate as the given GI/G/n

$$P_{M/M/n}(T_q > 0) = (n\mu - \lambda)W_{qM/M/n}$$

Then we replace the multiple servers in the GI/G/n queue with a single server having service rate

$$\mu' = \frac{\lambda}{P_{M/M/n}(T_q > 0)}.$$

Our assumption is that for a GI/G/n queue this approximation will create a GI/G/1 queue having approximately the same probability of waiting. We then approximate $P(T_q > 0)$ for the resulting GI/G/1 queue by interpolation.

We assume that $P(T_q > 0)$ can be approximated as a function of coefficients of variation of inter-arrival time and service time c_a and c_s: $P(T_q > 0) = f(c_a, c_s)$. We estimate $f(c_a, c_s)$ by computing the plane that passes through three points surrounding (c_a, c_s) for which $P(T_q > 0)$ is known or can be closely approximated. Expressed in the form $(c_a, c_s, f(c_a, c_s))$, these three surrounding points are: (0, 0, 0) $(1, c_s, f(1, c_s))$ and $(c_a, 1, f(c_a, 1))$ with $f(0,0) = 0$, $f(1, c_s) = \lambda/\mu'$, $f(c_a, 1) = r^k$. We can compute $k = 1/c_s^2$ to get Erlang distribution parameter k and r is the root of the characteristic equation: $\mu' r^{k+1} - (k\lambda + \mu')r + k\lambda = 0$.

Equation $f(0,0) = 0$ gives the probability of waiting for deterministic queues; Equation $f(1, c_s) = \lambda/\mu'$ is the probability of waiting for M/G/1 queues. In equation $f(c_a, 1) = r^k$ we are assuming that the inter-arrival time distribution can be approximated using an Erlang distribution and applying the formula for $E_k/M/1$ queues (Gross and Harris 2002).

The plane that passes though these points (0,0,0), $(1, c_s, \lambda/\mu')$ and $(c_a, 1, r^k)$ is given by

$$f(c_\alpha, c_s) = \frac{r^k c_s (1 - c_a) + (\lambda/\mu')(1 - c_s) c_a}{1 - c_s c_a}$$

This method for computing the probability of waiting is exact for M/M/n, M/G/1, and $E_k/M/1$.

COMPUTATIONAL ANALYSIS

To implement the approximation results, we develop a spreadsheet model (See figure 1) .The approximation results are compared with the Monte Carlo simulations. The errors are calculated by using

$$\%\ Error = 100\ \left|(\text{spread sheet}\ \sigma - \text{sim}\ \sigma)\right| / sim\ \sigma .$$

Different cases with the following combinations of parameters are compared:

Number of servers: 1, 2, 3, 10

Utilization: 0.4, 0.8, 0.9, 0.95, 0.99.

Coefficient of variation of inter-arrival times: 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5.

Coefficient of variation of service times: 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5.

Distribution of inter-arrival times: Normal, Gamma.

Distribution of service times: Normal, Gamma.

To evaluate the accuracy of our approximations, we conduct simulation experiments using the Extend simulation program. The testing of our approximations has been based on extensive simulation experiments.

Note Arrival rate and service rate must be in same units of time.
Arrival rate = 1/average interarrival time
Service rate = 1/average service time

COV(a) = standard deviation of interarrival times/average interarrival times.
COV(s) = standard deviation of service times/average service times.

Basic Inputs:	Number of Servers, S =	3
	Arrival Rate, λ =	27.00
	Service Rate Capacity of each server, μ =	10.00
	Coefficient of Variation of Inter-arrival time, COV(a) =	0.6
	Coefficient of Variation of Service time, COV(s) =	0.5

Inputs

The Waiting Line:	Average Number Waiting in Queue (Lq) =	2.26
	Average Waiting Time (Wq) =	0.08
	Probability of customer waiting P(t) =	0.68
	Probability of customer waiting no time P(0) =	0.32
	Standard deviation of waiting time σ =	0.11
Service:	Average Utilization of Servers (rho) =	0.90
	Average Number of Customers Receiving Service =	2.70
The Total System (waiting line plus customers being served):		
	Average Number in the System (L) WIP =	4.96
	Average Time in System (W) Cycle Time =	0.18
	Standard deviation of time in system σS =	0.12
	"Worst Case" Flow time = Average time in system + 3 σS =	0.55

Figure 1: Spreadsheet model to measure queuing performances

In this simulation research, we performed independent replications using 54000 minutes of simulation time and estimated 95% confidence intervals. Both Normal and Gamma distribution are used as general distribution. For Gamma distribution, when shape parameter k is positive integer, Gamma is reduced to Erlang. When k = 1, it is exponential. When k→∞, it is deterministic.

Simulation experiments confirm that the approximations perform remarkably well across a wide range of cases. In most of these cases, the standard deviation of the time in the system obtained with the spreadsheet was within 10% of that obtained in the simulation.

The exceptions are some of the cases in which the utilization was 0.99 or higher (high traffic queue) and/or the coefficients of variation of inter-arrival time or service time are 1.5 or greater. Notice that these are cases in which the performance of the queue itself becomes unstable. It is a situation in which a small change in a parameter would create a large change in the predicted performance. For example, for any queue the predicted performance is very sensitive to small changes in the parameters when the utilization is near 1. So even though the approximation does not work

very well in this situation, neither does simulation. When the parameters are restricted to ranges in which the queue is stable, our approximation works well for multiple server queues. It also it is insensitive to the distributions of the service time and inter-arrival times, when these distributions are "reasonable", i.e. normal or uniform. Heavy traffic queue approximation is beyond the scope of is paper, so the result of this paper is not applicable to heavy traffic queues.

Another limitation of the research is that the result is under the assumption that the coefficients of variation of the inter-arrival times and the service times are between 0 and 1.25, which is usual in practice. When coefficients of variation are greater than 1.5, the performance of the queue itself becomes very unstable. As noted by Whitt (1993), greater variability means less reliable approximation, because such descriptions evidently depend more critically on the missing information.

Thestandard deviation of the % error is 3.31 for all cases reported. As utilization increases, the % error standard derivation increases accordingly. For utilizations of 0.8, 0.9, and 0.95, they are 2.92, 3.16, and 3.96 respectively. As the number of server increases, the error standard derivation decreases, for instance, for number of servers 1, 2, 3, and 10, they are 3.67, 3.63, 3.55, and 2.31. Below we present selected examples comparing the approximation with simulation values. The data are grouped in different ways to show the effect of different parameters on the accuracy of the approximation.

5.1.Example of Results with Single Server Queues: The Impact of the Coefficients of Variation

Table 1 below demonstrates the accuracy of the approximation for single server queues over a range of values of the coefficients of variation for the inter-arrival times and the service times. In this set of problems, the arrival rate and service, rates were assumed to be $\lambda = 9$ and $\mu = 10$, giving a utilization of 0.9. Table 1 Comparison of approximations with simulation estimates of the standard deviation of waiting time in GI/G/1 for different coefficients of variation

The simulation results in the table are obtained using gamma distributions for the inter-arrival times and the service times.Thus the approximation of $E[s^3]$ was exact for these cases. The row and columns in bold represent cases for which the approximation formula is known to be exact. For these cases, the differences between the standard deviations given by the approximation and the simulation results are due to the sampling error in the simulation.

Table 1 show that the approximation methods work well for GI/G/1 queues for different combinations of

Table 1

Utilization = 0.9				*Service process*		
Ari proces	*method*	*CVs = 0.00*	*CVs = 0.25*	*CVs = 0.50*	*CVs = 0.75*	*CVs = 1.00*
CVa = 0.00	App	0.00	0.05	0.17	0.34	0.50
	Sim	0.00	0.04	0.16	0.31	0.51
CVa = 0.25	App	0.05	0.08	0.19	0.36	0.53
	Sim	0.04	0.07	0.17	0.34	0.54
CVa = 0.50	App	0.15	0.18	0.29	0.45	0.62
	Sim	0.13	0.16	0.27	0.39	0.62
CVa = 0.75	App	0.30	0.33	0.43	0.60	0.78
	Sim	0.28	0.31	0.41	0.55	0.78
CVa = 1.00	App	0.48	0.51	0.61	0.77	0.99
	Sim	0.48	0.53	0.63	0.79	1.01
CVa = 1.25	App	0.75	0.79	0.89	1.05	1.27
	Sim	0.73	0.84	0.85	1.08	1.27

coefficients of variation. The cases shown in table 1 represent fairly congested queues. For example, in the M/M/1 case, the average time in the system and the standard deviation of the time in the system, are each 1.00. These are ten times the average service time of 0.1 and the standard deviation of the service time which is also 0.1.

5.2.Example Results with Multiple Server Queues

In table 2, we show some results with multiple server queues. In these cases the service rate for each server is $\mu = 10$ and the arrival rate $\lambda = 9n$ giving a utilization of 0.9. Again the simulations for comparisons used gamma distributions for the inter-arrival time and service time distributions. We have bolded the cases for which our method is known to give the exact result.

Table 2 Comparisons of approximations with simulation estimates of the standard deviation of waiting time in GI/G/n queue for different number of servers for a fixed level of utilization and coefficients of variation, the approximation actually improves as the number of servers increase. However holding these parameters fixed and increasing the number of servers reduces congestion and the variability of the time in the system. One could argue that these are not valid comparisons.

Another way to compare would be to look at cases in the table, for which the standard deviation are about the same, (e.g. find the 1, 2, 3 server queues that have a standard deviation of about .5). When we do this we observe that the size of the error increases slightly as the number of servers increase.

5.3.Example Results with Normal Distributions

Table 3 below shows results of comparing the approximation method with simulation in which a normal distribution was used (instead of gamma) for the inter-arrival time and service time distributions. The error of the approximation is not much larger when the normal assumptions are used in the simulation.

Table 2

Utilization 0.9		*1 server*		*2 servers*		*3 servers*	
CVa	*CVs*	*Sim.*	*App.*	*Sim.*	*App.*	*Sim.*	*App.*
0	0	0	0	0	0	0	0
0	0.5	0.16	0.17	0.09	0.08	0.07	0.06
0	1	0.51	0.50	0.27	0.25	0.17	0.16
0	1.25	0.83	0.78	0.43	0.39	0.28	0.26
0.5	0	0.13	0.15	0.07	0.07	0.04	0.05
0.5	0.5	0.27	0.29	0.15	0.14	0.10	0.09
0.5	1	0.63	0.62	0.33	0.31	0.23	0.21
0.5	1.25	0.88	0.91	0.46	0.45	0.31	0.30
1	0	0.48	0.48	0.24	0.24	0.16	0.16
1	0.5	0.63	0.61	0.30	0.30	0.22	0.20
1	1	1.01	0.99	0.52	0.49	0.32	0.33
1	1.25	1.28	1.28	0.66	0.64	0.41	0.43
1.25	0	0.73	0.75	0.36	0.37	0.26	0.24
1.25	0.5	0.84	0.89	0.41	0.44	0.27	0.29
1.25	1	1.27	1.27	0.60	0.63	0.40	0.42
1.25	1.25	1.55	1.56	0.73	0.78	0.51	0.52

Table3 Comparisons of approximations with simulation estimates of the standard deviation of waiting time in GI/G/n queue for different simulation distributions

Table 3: Shows different distributions of arrival process and service process

Utilization = 0.9	*Methods*		
CVa = 0, CVs = 0	*Sim.(Normal)*	*Sim.(Gamma)*	*Approximation*
1 server	0	0	0
2 servers	0	0	0
3 servers	0	0	0
CVa = 0, CVs = 0.5	Sim. (Normal)	Sim. (Gamma)	Approximation
1 server	0.14	0.16	0.17
2 servers	0.08	0.09	0.08
3 servers	0.09	0.07	0.06
CVa = 0, CVs = 1	Sim.(Normal)	Sim.(Gamma)	Approximation
1 server	0.52	0.51	0.50
2 servers	0.30	0.27	0.25
3 servers	0.19	0.17	0.16
CVa = 0.5, CVs = 0	Sim.(Normal)	Sim.(Gamma)	Approximation
1 server	0.13	0.13	0.15
2 servers	0.07	0.07	0.07
3 servers	0.05	0.04	0.05
CVa = 0.5, CV = 0.5	Sim.(Normal)	Sim.(Gamma)	Approximation
1 server	0.26	0.27	0.29
2 servers	0.14	0.15	0.14
3 servers	0.08	0.10	0.09
CVa = 0.5, CVs = 1	Sim.(Normal)	Sim.(Gamma)	Approximation
1 server	0.64	0.63	0.62
2 servers	0.36	0.33	0.31
3 servers	0.27	0.23	0.21
CVa = 1, CVs = 0	Sim.(Normal)	Sim.(Gamma)	Approximation
1 server	0.49	0.48	0.48
2 servers	0.26	0.24	0.24

The simulation tests show that the standard deviation does not change dramatically when normal distributions are used instead of gamma distribution. This suggests that the standard deviation tends to be insensitive to "reasonable" changes in the distribution assumptions, and hence the approximation will work well for these different distributional assumptions.

CONCLUDING REMARKS

The role of time in healthcare delivery is becoming more important as the expectations for convenience and quality continue to rise. Patients are expecting increased availability of appointments and resources, shorter waits in treatment facilities, and quicker turnaround of results.In practical applications, the problems of operations management in health care service systems have recently attracted a lot of attention.Management scientists use techniques such as queuing theory and discrete event simulation to propose various appointment strategies under different clinics' settings.

Currently, these practices have no guidelines or frameworks to help identify an appropriate balance between physician capacity and patient panel sizes that are consistent with manageable patient backlogs. This research presents queuing models that we believe will be very helpful in this regard. In particular, these are the first spreadsheet models to explicitly estimating the variance of waiting time. The goal of this research 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 research provides 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.

As we have demonstrated, the modeling assumptions are that the first and second moments of the inter-arrival and service time distributions are known. Equation (1) is exact for G/M/n and M/G/1queues. Thus, the method for computing the coefficient of variation of waiting time in the queue is exact for any subset of these queues for which the exact probability of waiting and the second and third moments of the service time distribution is known. When the parameters are restricted to ranges in which the queue is stable our approximation works well for multiple server queues. As noted by Whitt (1993) and Kleinrock (1976), greater variability means less reliable approximation.

In the implementation, we assumed the service time distribution was gamma. The method for computing

the probability of waiting is exact for M/G/1, M/M/n.Thus; the method gives the exact coefficient of variation of waiting times for M/M/n and $E_k/M/1$, queues. While no model is a perfect representation of reality, we believe that these are useful for patient appointment system. Specifically, our results indicate that the spreadsheet model developed here is extremely reliable when patients take the available appointment.

When using the spreadsheet models to measure the variance of customer waiting time, we assume that the mean and the standard deviation of customer inter-arrival and service time distributions are known. Thus, all descriptions of the models depend only on the basic parameter 5-tuple: arrival rate λ, service rate μ, coefficient of inter-arrival time c_a, coefficient of service time c_s, and the number of server s. In practice, all these parameters are measurable for any health care service operations so the models have many potential applications. For instance, to implement appointment scheduling, we can measure the 5–tuple parameters and input the spreadsheet models to calculate the mean and standard deviation of waiting time. That is, we can measure patient's arrival rate λ(phone call rate), coefficient of patient's inter-arrival time c_a doctor's service rate μ, coefficient of doctor's service time c_s, and the number of doctors s. By using different λ, μ, c_a, c_s, and s, managers can estimate the worst case, such as mean waiting time $+3\ \sigma_q$, assuming the standard deviation of waiting time is normal distribution.

By using the spreadsheet and analyzing the case, managers can not only measure process flows and mean waiting time, but can also estimate the variance of waiting time to implement waiting time guarantee strategy. In addition, managers can also gain some important insights in health care management if conducting what-if analysis by inputting different parameters of coefficient of variations of arrival times and service times: (1) Variability causes loss of flow rate and effective capacity (2) Variability causes delays and congestions (3) In highly variable systems, waiting time increases nonlinearly with utilization (4) Capacity and variability reduction are substitutes in providing customer service. Process modeling helps managers understand real-world processes in detail and provides insights to the interaction among decisions about elements of service-delivery processes. They learn that easy, inexpensive changes can greatly improve turnaround time, and that the obvious process improvements do not always produce desirable results.

There are other healthcare areas where management science techniques will be useful, such as reducing delay in healthcare delivery, smoothing of elective admissions to reduce peak bed occupancy, and optimal deployment of ambulances. Quantitative techniques and data can help to present objective argument. Expert opinions could then be used to fine-tune the quantitative models. In our view, there is room for more management sciences to be applied in the healthcare settings. We hope this paper raises the awareness and adoption of management science applications among healthcare managers.

BIOGRAPHICAL NOTES

Xiaofeng Zhao is associate professor of operations and management science at College of Business, University of Mary Washington. He received his MBA degree from Indiana University of Pennsylvania and Ph.D. degree from University of Tennessee. Zhao's current research interests focus on service operations, supply chain risk management, lean/agile operations, and queuing theory applications.

Kenneth Gilbert is Ralph and Janet Heath Professor at College of Business Administration, University of Tennessee at Knoxville. Dr. Gilbert holds a B.S. degree in Mathematics from Berea College, an M.S. in Mathematic s from the University of Tennessee, and a Ph.D. in Management Science from the University of Tennessee. Dr. Gilbert regularly teaches a doctoral seminar in supply chain dynamics and also teaches in the UT Executive MBA program, in Manufacturing Management and is coordinator of the undergraduate curriculum in Lean Production. He also teaches in the UT Center of Executive Education Supply Chain Specialist certification course. He has served as a consultant to numerous companies.

References

[1] 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*, Vol. 36, No. 6, pp. 569-579.

[2] Albin,S.L., Barrett, J., Ito, D. and Mueller, J.E. (1990). "A queuing network analysis of a health center". *Queuing Systems*, Vol. 7, No. 1, pp. 51-61.

[3] Atkinson, J B. (2009). "Two new heuristics for the *GI/G/n*/0 queuing loss system with examples based on the two phase coaxial distribution". *The Journal of the Operational Research Society*, Vol. 60, No. 6, pp. 818-830.

[4] Bailey, N. T. J. (1952), "A Study of Queues and Appointment Systems in Hospital Outpatient Departments". *Journal of the Royal Statistical Society*, Series B Vol. 14, No. 2, pp. 185-99.

[5] Bennett, J. C., and Worthington, D. J. (1998). "An Example of a Good but Partially Successful OR Engagement: Improving Outpatient Clinic Operations", *Interfaces*, Vol. 28, No. 5, pp. 56-69.

[6] Bertsmas, D. (1990)."An analytic approach to a general class of G/G/s queuing systems". *Operations Research*, Vol. 38, No. 1, pp. 139.

[7] Cox, T, Birch all, J., and Wong, H. (1985). "Optimizing the Queuing System for an Ear, Nose, and Throat Outpatient Clinic". *Journal of Applied Statistics*, Vol. 12, No. 2, pp. 113-126.

[8] Fomudam, S., Herrmnn, J. (2007). "A Survey of Queuing Theory Applications in Healthcare", ISR Technical Report 2007, No.24, University of Maryland.

[9] Green, L. (2006a). "Queuing analysis in healthcare, in Patient Flow: Reducing Delay in Healthcare Delivery". Hall, R.W., ed., Springer, New York, Vol. 2, pp. 281-308.

[10] Green, L., Savin, S. (2008). "Reducing Delays for Medical Appointments: A Queuing Approach". Operations Research, Vol. 56, No. 6, pp. 1526–1538.

[11] Green, L. and Nguyen, V. (2001). "Strategies for Cutting Hospital Beds: The Impact on Patient Service". *Health Services Research*, Vol. 36, No. 2, pp. 421-42.

[12] Griffiths, J.D., Lloyd, N.P., Smithies, M. & Williams, J. (2006). "A queuing models of activities in an intensive care unit". *IMA Journal of Management Mathematics*, Vol. 17, No. 3, pp. 277-288.

[13] Gross, D., Harris, C. M. (2002). "Fundamentals of Queuing Theory". 2nd edition. John Wiley, New York.

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

[15] Kao, E.P.C. and Tung, G.G. (1981), "Bed allocation in a public health care delivery system". *Management Science*, Vol. 27, No. 5, pp. 507-520.

[16] Kimura, T. (1986). "A two-moment approximation for the mean waiting time in the GI/G/s queue". *Management Science*, Vol. 32, No. 6, pp. 751-763

[17] Kleinrock, L. (1976). "Queuing Systems", Volume I & II: Theory. John Wiley and Sons.

[18] Larson, R. C. (1987). "Perspectives on Queues: Social Justice and the Psychology of Queuing". *Operations Research*, Vol. 35, No. 6, pp. 895-905.

[19] McClain, J. (1976). "Bed planning using queuing theory models of hospital occupancy: a sensitivity analysis". *Inquiry*, Vol. 13, No. 2, pp. 167-176.

[20] Metters, K and Pullman, M. (2003). "Successful Service Operations Management". 2nd edition. Thomson-South Western.

[21] Noon, C., Hankins, C and Cote, M. (2003). "Understanding the Impact of Variation in the Delivery of Healthcare Services", *Journal of Healthcare Management*, Vol. 48, No. 2, pp. 82-97.

[22] Nosek, R. and Wilson, J. (2001). "Queuing theory and customer satisfaction: a review of terminology, trends, and applications to pharmacy practice". Hospital Pharmacy, Vol. 36, No. 3, pp. 275-279.

[23] Preater, J. (2002). "'Queues in health". Health Care Management Science, Vol. 5, No. 4, pp. 283-291.

[24] Sakasegawa, H. (1977). "An approximate formula Lq $= \alpha\beta\ \rho\ /\ (1 - \rho)$". Ann. Inst. Statist. Math. Vol. 29, pp. 67-75.

[25] Seelen, L.P. and TIJMS. (1984). "Approximations for the conditional waiting times in GI/G/c queue", Operations Research letters, Vol. 3, pp. 183-190.

[26] Shore, H. (1988). "Simple approximations for the GI/G/c queue-I: The steady-state probabilities". *The Journal of the Operational Research Society*, Vol. 39, No. 3, pp. 279-284.

[27] Trucker, J.B., Barone, J.E., Cecere, J., Blabey, R.G. and RHA, C. (1999). "Using queuing theory to determine operating room staffing needs". *Journal of Trauma*, Vol. 46, No. 1, pp. 71-79.

[28] Welch, J. D. (1964). "Appointment Systems in Hospital Outpatient Departments". *Operational Research Quarterly*, Vol. 15, No. 3, pp. 224-32.

[29] Whitt, W. (1993), "Approximations for the GI/G/m queue. Production and Operations Management", Vol. 2, No. 2, pp. 114-161.

[30] Whitt, W. (2004). "A diffusion approximation for the GI/G/n/m queue", Operations Research, Vol. 52, No. 6, pp. 922-941.

[31] Worthington, D. (1991). "Hospital waiting list management models". *The Journal of the Operational Research Society*, Vol. 42, No. 10, pp. 833-843.