The Modeling of Interrupts and Unplanned Absences in Health Care Operations

Stefan Creemers
IESEG School of Management, Lille, France
S.Creemers@ieseg.fr

Marc Lambrecht
Faculty of Business and Economics,
Research Center for Operations Management,
K.U.Leuven, Belgium
Marc.Lambrecht@econ.kuleuven.be

The health care sector is a fast-growing segment of GNP in almost every economy. No wonder that we witnessed a tremendous increase in research to improve both medical practice and management practice. Patient flow management is an example of such a management practice and represents the ability of the health care system to serve patients quickly, reliably, and efficiently as they move through stages of care. This is basically a supply chain issue in the service sector in which we try to match demand and supply. Unfortunately patients still experience delays and waiting lists. This is an immediate consequence of a mismatch between demand and supply. A queueing model offers an excellent tool to analyze and improve the performance of health care systems. The purpose of this article is to discuss differences with the modeling of manufacturing systems (or more broadly supply chain issues in the health care industry compared to supply chain issues in services) and to focus on modeling issues in patient flow. In this work, we concentrate on service outages and develop new expressions to assess their impact on waiting lists and delays. Using data obtained from a Belgian hospital, these expressions are evaluated through a number of queueing models. We provide a comparison of the different queueing models and indicate which queueing technique is most appropriate to model a health care system.

Keywords: service outages, health care, queueing model, case study, health care supply chain, delays

Introduction

An important feature of health care processes (or services in general) is that the demand for resources is to a large extent unscheduled. As a consequence, there is a permanent mismatch between the demand for a treatment and the available capacity. This phenomenon can be described as a generic supply chain problem. Moreover, timely care is very important in health care processes (the sense of urgency is almost always present). No wonder that health care is riddled with delays. There is no need to come up with a convincing example; we have all experienced that phenomenon. Delays are highly undesirable, not only from a psychological point of view (patient satisfaction) but also from an economic point of view. Government reimbursement systems are increasingly based on a justified length of stay (JLoS) system. DRGs (diagnosis related groups) are characterized by a minimum and maximum length of stay (depending on parameters such as severity of the illness, age of the patient, etc.). If a patient is dismissed before the JLoS, the hospital still collects a full reimbursement. However, if the

ISSN print 1625-8312
ISSN online1624-6039

patient remains in care for a period that exceeds the limit of the JLoS, the hospital has to pay for the extra costs involved. The JLoS of a DRG is determined in conjunction with a national average length of stay. The system stimulates hospitals to continuously improve their performance. Moreover, improper scheduling and malfunctioning logistical systems cause length of stays that are too long. Insurance companies may reject reimbursement of these “denied days” because the delay is not medically necessary (Hall, 2006a). Delays also create a “hidden” hospital in analogy with the hidden company. In other words, such a hospital creates wasteful overhead. It is clear that the language we use here can be perfectly translated into an industrial environment in which terms such as *customer service*, *backlogs*, and *lead times* are commonly used. Throughout this article we will use supply chain terminology to characterize the health care environment.

Randolph Hall (2006a, 2006b) coined the term *patient flow*. It represents the ability of the health care system to serve patients quickly, reliably, and efficiently as they move through stages of care. Queue and delay analysis can produce dramatic improvements in medical performance, patient satisfaction, and cost efficiency of health care. Health care systems can be represented as a complex queueing network. The queueing models are helpful to determine the capacity levels (and the allocation of capacity) needed to respond to demands in a timely fashion (minimizing the delay). There is a demand side (the patient mix and the associated variability in the arrival stream) and a supply side (the hospital resources such as surgeons, nurses, operating rooms, waiting rooms, recovery areas, imaging machines, and laboratories) in any health care process. Moreover, both demand and supply are inherently stochastic. This stochastic nature creates disturbances and outages during the process. It is the combination of capacity analysis and variability that makes queueing theory so attractive. The major objective is to identify factors influencing the flow time of patients, to identify levers of improvement, and to analyze trade-offs.

Health care systems, however, have a number of specific features making the modeling much more difficult than a typical industrial manufacturing process. These features pose important methodological challenges. This is the subject of section 2. In section 3 we focus on one of these features, namely service outages, and develop new expressions to assess the impact on patient flow time. Section 4 compares a number of queueing models (both parametric decomposition approaches as well as Brownian queueing models are addressed) and provides insight into which approach is best to model health care systems. Section 5 draws some conclusions.

> An important feature of health care processes is that the demand for resources is to a large extent unscheduled

Using queueing models to reduce delays in a health care system

Queueing models have been applied in numerous industrial settings and service industries. The number of applications in health care, however, is relatively small. This is probably due to a number of unique health care-related features that make queueing problems particularly difficult to solve. In this section, we will review these features and when appropriate we will discuss briefly the methodological impact.

Before we dig into this issue, let's first discuss two important modeling issues in health care: the performance measures and the issue of pooled capacity.

The performance measures in health care systems focus on internal and external delays. The internal delay refers to the sojourn time of patients inside the hospital before treatment. The external delay refers to the phenomenon of waiting lists. Manufacturing systems may buffer with finished goods inventory, but service systems rely more on time buffers and capacity buffers. Another important performance measure is related to the target occupancy (utilization) levels of resources. Average occupancy targets are often preferred by government and other institutional agents. Therefore, higher occupancy levels are preferred, but this results in longer delays. We are often confronted with conflicting objectives. Instead of determining capacity needs based on (target) occupancy levels, it is preferable to focus on delays. The key issue in delay has to do with the tail probability of the waiting time. The tail probability refers to the probability that a patient has to wait more than a specified time interval. Capacity needs (e.g., staffing) of an emergency department should be based on an upper bound on the fraction of patients who experience a delay of more than a specific time interval before receiving care from a physician (Green & Soares, 2007).

The second modeling issue has to do with capacity pooling. In general, pooling refers to the phenomenon that available inventory or capacity is shared among various sources of demand (well-known examples are location pooling, commonality, or flexible capacity). Pooling is based on the principle of aggregation and mostly comes down to the fact that we can handle uncertainty with less inventory or capacity. In health care systems, resources are usually dedicated to specific patient types, hospitals have separate units or departments by diagnostic type, and bed flexibility is almost nonexistent. As a result, capacity pooling is absent. This explains the fact that most queueing models reported in the literature deal with parts of the hospital. Queueing models can be used to model hospitalwide systems and to evaluate the benefits of greater versus less specialization of care units or other resources (scanners, labs, etc.).

Let's now turn to a number of unique health care-related features making queueing models in health care difficult to model and to solve.

1. Time-varying demand

Queueing models usually assume time-independent (input) demand rates. Health care facilities generally experience different demands during a day, a week, or a season. Arrivals consist of acute (unscheduled) and elective (scheduled) patients. In other words, part of the input cannot be controlled and another part can be scheduled. As a consequence, staffing has to be adjusted constantly. The long-term steady-state probability distributions for queue length or delay are usually assumed to be independent of time. In health care systems we should rely more on time-varying arrival rates and time-varying server availability and time-dependent waiting times (Green & Soares, 2007; Ingolfsson et al., 2002). Green (2006) proposes a stationary independent period-by-period approach to determine how to vary staffing to meet changing demand.

2. Waiting creates additional work

Hall (2006b) points out that waiting creates additional work for clinicians because patients must be monitored. This situation does not occur in a manufacturing environment in which buffers typically do not consume resources.

3. Re-entry of patients and stochastic routings

During consultation, patients may be routed to different facilities. The routing of a patient through hospital facilities is not deterministic. Instead, during the diagnosis stage there is a probabilistic routing. Moreover, patients require in many cases several consultations before, for example, surgery. Even after a patient is discharged from the hospital after surgery and recovery, the patient is subjected to a number of follow-up consultations. In other words, the queueing model must take care of the re-entry of patients, creating additional work on top of dealing with new patients. In most cases, the re-entry is correlated.

4. Time blocks for consultation and surgery

In most queueing models time is considered as continuous and events are spread out over this continuous time scale. In services in general and in health care more specifically, resources are not continuously available. Instead, time is divided into time blocks for consultation (e.g., twice a week) or surgery (e.g., one day per week). Consequently we have to focus on service processes in which the service takes place during predefined service epochs. Vacation models observe the queueing behavior of such systems in which servers are available during certain time blocks and are on "vacation" during the other time intervals.

5. Capacity-related issues

Hospitals operate within strict business restrictions. Resources are usually very scarce and consequently hospitals operate under high-capacity utilization conditions. So-called heavy traffic conditions that assume that all stations in the network are critically loaded are present. In such an environment, traditional parametric decomposition approaches may not yield accurate results for the performance measures. Other approaches may be necessary such as Brownian queueing models.

6. Modeling of absences, disturbances, and interruptions

An important determinant of the flow time is variability. We distinguish two types of variability. Natural variability is variability that is inherent to the system process and is much more substantial in health care as compared to manufacturing environments. Second, we have variability that can be related or assigned to a specific external cause. This variability is caused by unplanned absences of medical staff or interruptions during service operations. It is well known that variability induces waiting time. As a result the time available during consultation is often exceeded. This in turn is remedied by allowing overtime. Unfortunately, overtime modeling is a nontrivial issue in queueing.

7. Queueing discipline

The first-in-first-out (FIFO) assumption is common in queueing. From the moment that other queueing disciplines are introduced, the model becomes very complex. Unfortunately, disciplines other than FIFO are quite common in health care (triage system for emergencies, priority changes for medical reasons, etc.). Moreover, queue lengths are often limited or we have a closed queueing system (patients are allowed to enter the system if another patient leaves the system). All of these features dramatically complicate the modeling exercise.

In what follows, we will touch on several of these aspects. However, we will mainly focus on the modeling of interrupts and unplanned absences. It is well known that patients also use nonclinical factors in their choice of hospitals. Important nonclinical factors are length of waiting lists, timeliness of the service, ease of making an appointment, waiting time during the service session, and so on. Most of these inconveniences are caused by service variability. That's what we model in the next section.

The modeling of interrupts and unplanned absences

With respect to service outages in health care, a large body of literature exists. Outages in a hospital setting have been the subject of discussion in Babes and Sarma (1991), Liu and Liu (1998), and Chisholm et al. (2000, 2001) among others. There is a consensus on the harmful effects of outages on patient flow times as well as on the quality of service. Outages result in congestion, unstable schedules, and, most important, overtime for staff members. We refer to Easton and Goodale (2005) for an excellent treatment of this issue.

In this section, we focus on unplanned absences of medical staff and interruptions during service operations, which have a major impact on flow times. Doctors and medical staff face various obligations that they have to attend to (making morning rounds, answering phones, patient check-ups, daily management, etc.). In addition doctors often combine a hospital job and private consultation. These phenomena may cause a variable arrival pattern at the hospital (Liu & Liu, 1998) and may lead to interruptions during the treatment process (Chisholm et al., 2000, 2001; Easton & Goodale, 2005). It is clear that hospital environments are characterized by substantial amounts of variability. As is argued in the literature (Hopp & Spearman, 2000), variability induces waiting times. Because in service industries variability cannot be countered by means of inventory in the traditional sense, in health care patients will have to wait until capacity becomes available (Vissers et al., 2001; Vandaele & De Boeck, 2003; Sethuraman & Tirupati, 2005). In addition to the time buffer, hospitals often have to rely on a capacity buffer to mitigate the impact of variability and to maintain required service levels.

In order to model service processes liable to outages, queueing theory proves to be an ideal tool. With respect to service outages and server unreliability, we face a vast amount of queueing literature. Surveys on the machine interference problem and server unreliability may be found in Stecke and Aronson (1985) and Haque and Armstrong (2007). Unreliable servers are often modeled using vacation models. In the past, queueing systems with server vacations have received a lot of attention in the queueing literature. Vacation models observe the queueing behavior of systems in which the server begins a vacation (i.e., becomes unavailable) when certain conditions are met. For instance, a doctor's office has opening hours on Tuesday afternoons and on Friday evenings. On Tuesday, after service completion of the last patient, the doctor leaves on a "vacation" until Friday evening at which time service is resumed. At the end of service on Friday, a vacation is initiated until next Tuesday afternoon. We illustrate this process in Figure 1.

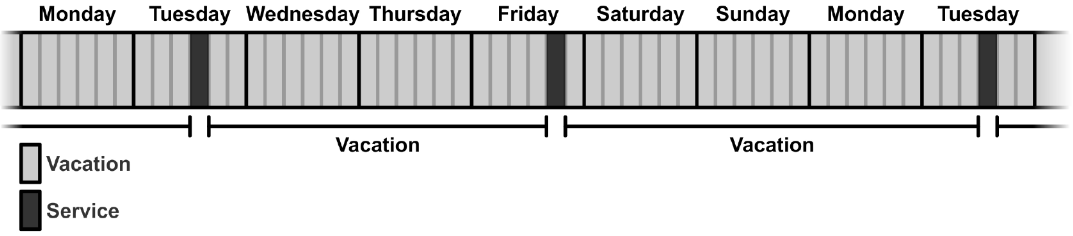
Next to the modeling of planned absences (e.g., a working schedule), vacation models may also be used to model unplanned server interruptions (e.g., a doctor who is called away for an emergency). A wide variety of vacation models exists. For a general overview, we refer to Doshi (1986) and Takagi (1988). A more recent yet less general survey can be found in Vishnevskii and Semenova (2006).

In this work, however, we do not focus on vacation models. Instead, we consider an alternative, more intuitive approach to model service outages. This approach was first suggested by Hopp and Spearman (2000), who proposed a transformation of the service process times to account for service outages. The results of Hopp and Spearman (2000) are widely accepted in the literature (see for instance Lambrecht et al., 1998) and have been further developed by Creemers and Lambrecht (2007). In this work, we summarize the most important results on the subject. In what follows, we first discuss the difference between preemptive and nonpreemptive outages. Then, we provide the means to model them.

A. Outages, classification, and impact

As was indicated previously, the service process of a patient may be interrupted or postponed. These outages will increase the natural service times (i.e., the raw service time excluding any impact resulting from outages). We call these increased, adjusted service times *effective service times*. It is the total time "seen" or "experienced" by a patient at a workstation. The effective process time random variable is of primary interest to determine flow times.

Figure 1
Illustration of a Vacation Model.



We distinguish between preemptive and nonpreemptive outages. Preemptive and nonpreemptive outages will affect the service process and will give rise to increased levels of traffic intensity (resulting in the so-called effective utilization rate or effective traffic intensity).

Let us first focus on nonpreemptive outages. Nonpreemptive outages typically occur between jobs rather than during jobs. They occur at the beginning of each service epoch (e.g., at the start of a consultation work shift) whenever a doctor or another member of the medical staff is absent (e.g., due to late arrival). We may refer to such an outage as an unplanned absence and define the mean and variance of the amount of time absent as T and s^2_T, respectively (i.e., absence times are allowed to follow a general distribution). Furthermore, we assume an average number of patients (represented by n) arrive in between two consecutive absences. This is an important feature of the model. Indeed, n may be considered as the number of patients in a service time block (e.g., a consultation work shift). Each start of a time block may induce a delay due to an absence. In other words, the number of patients in a time block is a decision variable and is comparable to a lot-sizing decision. Evaluating different time block sizes (i.e., different values of n) may provide key managerial insights.

Next to nonpreemptive outages, we also allow for preemptive outages to take place. Preemptive outages occur whenever a doctor is interrupted during a consultation activity. These interruptions will be modeled in an approach that builds on the tradition set by Hopp and Spearman (2000). They are characterized by a mean time to interrupt (τ_i) and a mean time to resolve (τ_r). The model presented in Hopp and Spearman (2000) presumes interrupts occur only during actual service time. However, in a hospital setting, it is possible that interrupts take place during the resolve time induced by a previous interrupt as well. For instance, if the service process of a patient is interrupted by a phone call, it is still possible for a doctor to be called away for an emergency, to receive another call, and so on.

In what follows, we present the main results on nonpreemptive as well as preemptive outages. In a final subsection, we present results on the joint occurrence of nonpreemptive and preemptive outages.

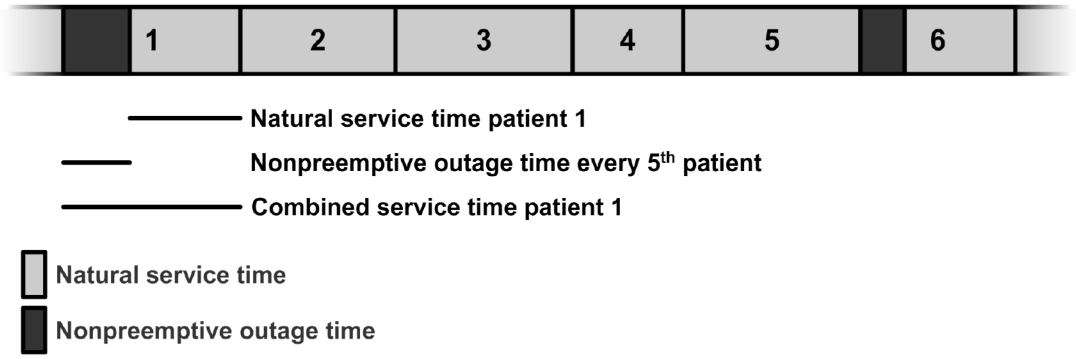
B. Nonpreemptive outages

We define a nonpreemptive outage to occur whenever the succession of two events is based on the number of services performed in between (hence, set-ups, rework, maintenance, etc. are all extensions that are able to capitalize on the technique discussed in this section). Applied to our setting, we assume that n patients are treated (on average) in between two consecutive absence possibilities. Assume that the length of services and absence times do not depend on the service history (i.e., they are independent of prior service and absence times). The absence times themselves are distributed following a probability density function f_T (x). The average absence time and its variance are represented by T and s^2_T, respectively. The service time of the n^{th} patient includes part service time, part absent time. We refer to the service time of the n^{th} patient as the combined service time. We illustrate these concepts in Figure 2.

The probability density function of the combined service times equals $f_c\,(x + y) = f\,(x)\,f_T\,(y)$, where f(x) is the probability density function of the natural service times with mean $\overline{X}$ and variance s^2_X . One can consider the services that are preceded by an absent period as a separate class of patients that has a probability $1/n$ of randomly being picked in front of the workstation (e.g., the doctor's office). The other services as a whole have a probability $(n - 1)/n$ of randomly being picked. Therefore, we can define the mean service times including the effect of absence times as follows:

$$\begin{aligned}\overline{X}_s &= \left[\left(\frac{n-1}{n}\right)\int f(x)x\,dx\right]+\\ &\quad\left[\frac{1}{n}\iint f(x)f_T(y)(x+y)\,dy\,dx\right],\\ &= \overline{X}+\frac{T}{n}.\end{aligned}$$

Figure 2
Illustration of the Combined Service Time.


1
2
3
4
5
6
Natural service time patient 1
Nonpreemptive outage time every 5th patient
Combined service time patient 1
Natural service time
Nonpreemptive outage time

With respect to the variance of the service time (including absence times), we develop the following expression:

$$\begin{aligned} s_s^2 &= \left[\left(\frac{n-1}{n}\right)\int f(x)\left(x-\overline{X}_s\right)^2 dx\right]+ \\ &\quad \left[\frac{1}{n}\iint f(x)f_T(y)\left(x+y-\overline{X}_s\right)^2 dydx\right], \\ &= s_X^2+\frac{s_T^2}{n}+T^2\left(\frac{n-1}{n^2}\right). \end{aligned}$$

This expression is equivalent to that of Hopp and Spearman (2000) and is valid under the assumption that the combined service times as well as ordinary service times are independently distributed.

C. Preemptive outages

We refer to service interruptions as preemptive outages. Doctors being called away on emergencies, answering phone calls, and so on are typical examples. The average time between two consecutive interrupts is defined as τ_i whereas τ_r refers to the average time it takes to resolve an interruption. Preemptive outages prove to be more difficult to model because they occur after the elapsing of a variable amount of time (i.e., a mean time to interrupt τ_i) rather than after a number of patients are processed. Under the assumption that the time between two consecutive interrupts is exponentially distributed, exact expressions for mean and variance have been obtained.

With respect to preemptive outages, we make a distinction between two different scenarios. On the one hand, one might presume preemptive outages occur only during actual service time. As such, preemptive outages do not take place during the resolve times induced by previous outages. Notice that this does not imply that the service process of a single patient cannot be interrupted more than once. On the other hand, one might assume preemptive outages to occur during resolve times as well (e.g., as indicated previously, doctors may be interrupted when already engaged in resolving a previous interrupt). Although this latter instance can be seen as an extension of the former, we will first discuss outages occurring exclusively during actual service time. Define $\tau_{r_o}(j)$ as the resolve time of the j^{th} preemptive outage that occurred during the service process of one and the same patient. The mean and variance of the resolve times are given by τ_r and s^2_r. In addition, resolve times of different outages are assumed to be independent and identically distributed (i.i.d). The service process of a patient thus faces the probability of encompassing several interrupts that prolong its service duration. The service time of a patient (including interrupts) can be expressed as

$$\overline{X}_i=\overline{X}+\sum_{j=1}^{J_0}\tau_{r_0}(j).$$

As such, the average service time $\overline{X}_i$ incorporates both the natural service time X as well as the resolve times of interrupts that occurred during service. Moreover, J_0 denotes the number of preemptive outages that occurred during the service process of a unit. J_0 is a random variable that follows a Poisson distribution (i.e., we assume the time between two consecutive interrupts to be exponentially distributed). Hence, its mean and variance both equal $\overline{X}/\tau_i$ (i.e., the mean service time divided by the mean time for an interrupt to occur). We face a sum of random variables (the resolve times $\tau_{r_o}(j)$) in which the number of random variables (the number of interrupts J_0) is a random variable itself. Assume that J_0 and $\tau_{r_o}(j)$ $(\forall j\in\{0,1,\ldots\})$ are i.i.d. variables. In addition, assume the mean as well as the variance of $\tau_{r_o}(j)$ to be equal for all $j\in\{0,1,\ldots\}$. Therefore, the mean and variance of the sum of J_0 random variables $\tau_{r_o}(j)$ can be expressed as (Dudewicz & Mishra, 1988):

$$E[S_0]=E[J_0]E\left[\tau_{r_o}(j)\right],$$

$$s_{S_0}^2=E[J_0]s_r^2+E\left[\tau_{r_o}(j)\right]^2 s_{J_0}^2,$$

where S_0 is the random variable representing the sum of J_0 resolve times $\tau_{r_o}(j)$. In other words, we have that

$$S_0=\sum_{j=1}^{J_0}\tau_{r_0}(j).$$

The mean and variance of the sum of resolve times can be defined as

$$E[S_0]=\overline{X}\frac{\tau_r}{\tau_i},$$

$$s_{S_0}^2=\overline{X}\frac{s_r^2+\tau_r^2}{\tau_i}.$$

We can now express the mean service time including the effect of interrupts as follows:

$$\overline{X}_i=\overline{X}\frac{\tau_i+\tau_r}{\tau_i}.$$

This corresponds to the expression presented in Hopp and Spearman (2000) in which the natural service time is divided by an availability factor in order to incorporate the effect of interrupts. Next, we have a look at the variance of the service times including the effect of preemptive outages during service times. We start with the expression of the second moment:

$$E\left[X_i^2\right]=\left[\left(s_X^2+\overline{X}^2\right)\left(1+\frac{\tau_r}{\tau_i}\right)^2\right]+s_{S_0}^2.$$

Using this expression it is easy to obtain the variance of the service times including the effect of interrupts:

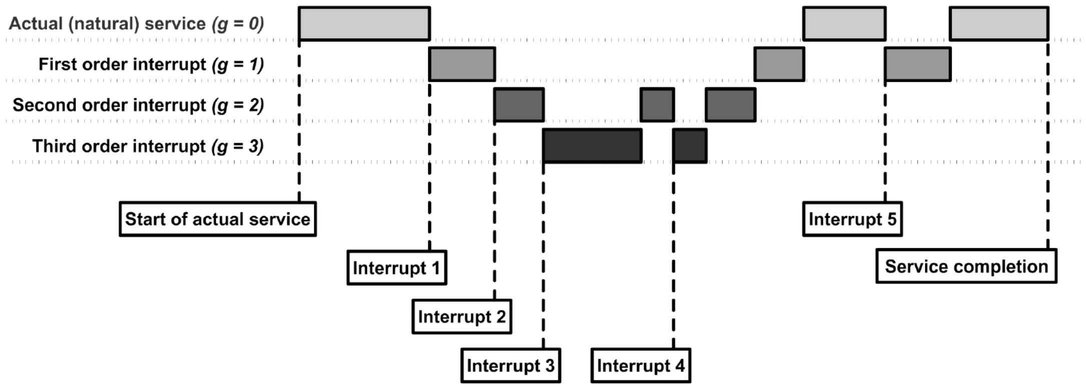
$$s_i^2=\left[s_X^2\left(1+\frac{\tau_r}{\tau_i}\right)^2\right]+s_{S_0}^2.$$

This expression once more matches the formula derived in Hopp and Spearman (2000). These expressions hold if and only if the Poisson-distributed preemptive outages take place during service itself. In what follows, we relax this assumption and allow for interrupts to take place during the resolve times induced by previous interrupts.

In order to approach this problem, we divide the interrupts into different sets. Let g (where g is an element of the set containing the natural numbers) denote the set index. We define $\tau_{r_g}(j)$ to be the resolve time of the j^{th} interrupt belonging to the set of index g (i.e., the interrupt is said to be of order g). Without loss of generality, assume that interrupts of order 0 occurred during actual service, interrupts of order 1 occurred during the resolve times of interrupts of order 0, and so on. In

Figure 3

Multilevel Interruptions During the Service Process of a Patient.


Actual (natural) service *(g = 0)*
First order interrupt *(g = 1)*
Second order interrupt *(g = 2)*
Third order interrupt *(g = 3)*
Start of actual service
Interrupt 1
Interrupt 2
Interrupt 3
Interrupt 4
Interrupt 5
Service completion

general, interrupts of order g took place during the resolving of interrupts of order (g - 1). Figure 3 provides further insight.

In addition, define S_g as the sum of resolve times corresponding to interrupts of order g. We have that

$$S_g = \sum_{j=1}^{J_g} \tau_{r_g}(j),$$

where J_g is the number of interrupts belonging to the set of index g. J_g follows a Poisson distribution and its mean and variance equal

$$E[J_g] = E[J_g^2] - E[J_g]^2 = \overline{X}\frac{1}{\tau_i}\left(\frac{\tau_r}{\tau_i}\right)^g.$$

One can infer that

$$E[S_g] = \overline{X}\frac{\tau_r}{\tau_i}\left(\frac{\tau_r}{\tau_i}\right)^g,$$

$$s_{S_g}^2 = \overline{X}\frac{s_r^2 + \tau_r^2}{\tau_i}\tau_r\left(\frac{\tau_r}{\tau_i}\right)^g.$$

Using the same reasoning as applied previously, one can express the mean service time including the effect of all order interrupts as follows:

$$\overline{X}_i = \overline{X}\frac{\tau_i}{\tau_i - \tau_r}.$$

Using these parameters, the second moment may be expressed as follows:

$$E\left[X_i^2\right] = \left\{\left(s_X^2 + \overline{X}^2\right)\left[1 + \frac{2\tau_r}{\tau_i - \tau_r} + \left(\frac{\tau_r}{\tau_i - \tau_r}\right)^2\right]\right\} + \overline{X}\left(\frac{s_r^2 + \tau_r^2}{\tau_i - \tau_r}\right).$$

As a result, the variance of the service time (including the impact of all order interrupts) is given by

$$s_i^2 = \frac{s_X^2\tau_i^2 + \overline{X}(\tau_i - \tau_r)(s_r^2 + \tau_r^2)}{(\tau_i - \tau_r)^2}.$$

D. Combining preemptive and nonpreemptive outages

In many hospital settings, both preemptive and nonpreemptive outages may surface. Although it is impossible to interrupt the service process in the instance of a nonpreemptive outage (e.g., a doctor who arrives late), we consider only the case in which both types of outages cannot occur simultaneously. The average service time incorporating this combined effect can be expressed as follows:

$$\begin{aligned}\overline{X}_{Ti} &= \left[\left(\frac{n-1}{n}\right)\int f_i(x)x\,dx\right] + \\ &\quad \left[\frac{1}{n}\iint f_i(x)f_T(y)(x+y)\,dy\,dx\right], \\ &= \overline{X}_i + \frac{T}{n},\end{aligned}$$

where f_i (x) is the probability density function of service times including the effect of all order interrupts. Its mean and variance are given by $\overline{X}_i$ and $s^2{}_i$, respectively. We refer to $\overline{X}_{Ti}$ as the effective service time while it equals the service time experienced by the patient (and as such includes the impact of outages). The variance of the effective service times at the consultation workstation may be expressed as

$$\begin{aligned}s_{Ti}^2 &= \left[\left(\frac{n-1}{n}\right)\int f_i(x)\left(x - \overline{X}_{Ti}\right)^2 dx\right] + \\ &\quad \left[\frac{1}{n}\iint f_i(x)f_T(y)\left(x + y - \overline{X}_{Ti}\right)^2 dy\,dx\right], \\ &= s_i^2 + \frac{s_T^2}{n} + T^2\left(\frac{n-1}{n^2}\right).\end{aligned}$$

These results allow us take service outages into account when assessing hospital performance measures.

In what follows we adopt these results to model the orthopedic department of a Belgian hospital.

Comparison of Different Queueing Models

The orthopedic department that is used as an illustrative case study may be represented as an open re-entry queueing network that consists of five *G/G/m* workstations (where *m* is the number of servers in each workstation). At these workstations, a FIFO service discipline is in force. The orthopedic department currently has six surgeons and can rely on a medical staff of more than 15 nurses. On a yearly basis 3,300 patients are being operated on and more than 13,000 patients receive consultations. The orthopedic department has three consultation rooms and claims the capacity-equivalent of two operating theatres. Empirical data gathered at the orthopedic department serves as the input of the queueing models.

Figure 4
Patient Flow at the Orthopedic Department of a Belgian Hospital.

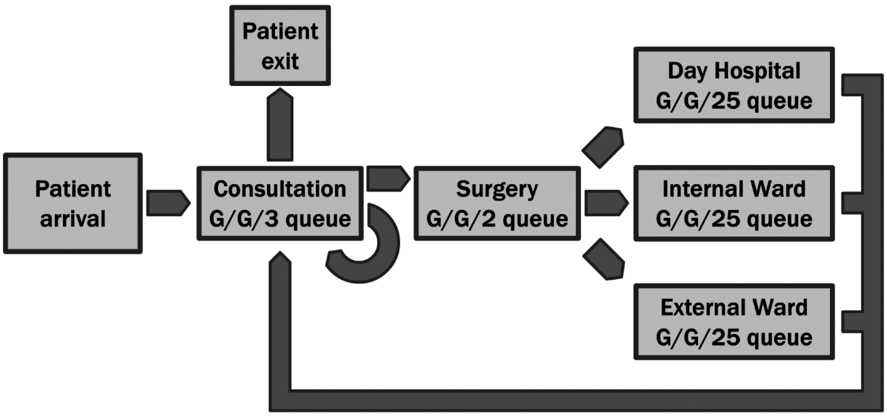

Patient exit
Patient arrival
Consultation G/G/3 queue
Surgery G/G/2 queue
Day Hospital G/G/25 queue
Internal Ward G/G/25 queue
External Ward G/G/25 queue

Table 1
Average Patient Waiting Time (in days) at the Different Workstations.

	Consultation	Surgery	Day Hospital	Internal Ward	External Ward
$E[W_{\text{Kingman}}]$	5.0589	3.9543	0.7971	5.2403	8.0969
$E[W_{\text{Whitt}}]$	5.0591	3.9530	0.7971	5.2033	8.0966
$E[W_{\text{Brownian}}]$	7.7226	5.4172	0.2792	1.1966	5.0012
$E[W_{\text{Simulation}}]$	5.401	3.4620	0.7971	5.1193	8.1013

The process of recovery takes place in an internal or external ward or in the day hospital. In each of these wards, a capacity of 25 beds is reserved for orthopedic patients. The treatment process of a patient starts with one or more consultations, followed by surgery, recovery, and a number of follow-up consultations. The flow of patients at the orthopedic department is depicted in Figure 4.

To model the patient flow at the orthopedic department, we use parametric decomposition approaches as well as a Brownian queueing model. With respect to the parametric decomposition approaches, we use the Kingman equation (Hopp & Spearman, 2000) and the approximation derived by Whitt (1993) to assess patient flow times. Simulation is used as a validation tool. We refer to Creemers and Lambrecht (2007) for a detailed discussion of the model and the data for the input parameters.

For each of the workstations, the average patient waiting times (corresponding to each of the queueing models) are presented in Table 1. When comparing the results using a simulation model, it is clear that the parametric decomposition approaches work best. With respect to the parametric decomposition approaches themselves, only minor differences are observed. Because the Whitt procedure is computationally much more demanding, the Kingman equation is the preferred flow time expression to assess patient waiting time at a complex hospital system.

It can easily be shown that interrupts have a detrimental impact on patient flow time. We can change the mean time to interrupt (e.g., during consultation) and calculate the impact on the expected patient waiting time. This impact is exponential for heavy traffic systems, that is, systems that operate under high workload, a situation often encountered in practice. There are practical guidelines for hospital decision makers to minimize the impact of interrupts on the service process, for example, filtering of nonurgent communication towards medical staff, pooling of paging of doctors, automation of administrative tasks, and the use of information systems.

Conclusion

Queueing models are very useful to quantify the relationship among capacity utilization, waiting time, and patient (customer) service. Increasing the capacity utilization, although financially attractive, may require unacceptably high patient delays and poor customer service. The presence of high degrees of variability (outages, absences, interruptions, etc.) negatively affects the performance of the health care system. In this article, we show that this capacity and variability analysis in a health care environment results in queueing models that are different from queueing models in an industrial setting. We developed new expressions to assess the impact of service outages typical in health care settings. To conclude, we compared a number of different queueing models and demonstrated that a simple parametric decomposition approach (i.e., the Kingman equation) provides the best performance when modeling complex hospital systems.

References

Babes, M., & Sarma, G. V. (1991). Outpatient queues at the Ibn-Rochd Health Center. *The Journal of the Operational Research Society*, 42(10), 845-855.

Chisholm, C. D., Collison, E. K., Nelson, D. R., & Cordell, W. H. (2000). Emergency department workplace interruptions: Are emergency physicians "interrupt-driven" and "multitasking"? *Academic Emergency Medicine*, 7(11), 1239-1243.

Chisholm, C. D., Dornfeld, A. M., Nelson, D. R., & Cordell, W. H. (2001). Work interrupted: A comparison of workplace interruptions in emergency departments and primary care offices. *Annals of Emergency Medicine,* 38(2), 146-151.

Creemers, S., & Lambrecht, M. R. (2007). *Modeling a healthcare system as a queueing network: The case of a Belgian hospital.* Research Report 0710. Leuven, Belgium: Department of Decision Sciences & Information Management, Research Center for Operations Management, Katholieke Universiteit.

Doshi, B. T. (1986). Queueing systems with vacations - A survey. *Queueing Systems: Theory and Applications,* 1(1), 29-66.

Dudewicz, E., & Mishra, S. (1988). *Modern mathematical statistics.* New York: John Wiley.

Easton, F. F., & Goodale, J. C. (2005). Schedule recovery: Unplanned absences in service operations. *Decision Sciences*, 36(3), 459-488.

Green, L. (2006). Queueing analysis in healthcare. *In R. Hall (Ed.), Patient flow: Reducing delay in healthcare delivery* (pp. 281-307). New York: Springer.

Green, L., & Soares, J. (2007). Computing time-dependent waiting time probabilities in M(t)/M/s(t) queueing systems. *M&SOM Manufacturing & Service Operations Management,* 9(1), 54-61.

Hall, R. (2006a). *Patient flow: Reducing delay in healthcare delivery*. New York: Springer.

Hall, R. (2006b). *2006 patient flow: The new queueing theory for healthcare*. OR/MS Today, 23(3), 36-40.

Haque, L., & Armstrong, M. J. (2007). A survey of the machine interference problem. *European Journal of Operational Research*, 179, 469-482.

Hopp, W., & Spearman, M. (2000). *Factory physics: Foundations of manufacturing management*. New York: Irwin/McGraw-Hill.

Ingolfsson, A., Haque, A., & Umnikov, A. (2002). Accounting for time-varying queueing effects in workforce scheduling. *European Journal of Operational Research*, 139, 585-597.

Lambrecht, M. R., Ivens, P. L., & Vandaele, N. J. (1998). Aclips: A capacity and lead time integrated procedure for scheduling. *Management Science*, 44(11), 1548-1561.

Liu, L., & Liu, X. (1998). Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society,* 49(12), 1254-1259.

Sethuraman, K., & Tirupati, D. (2005). Evidence of bullwhip effect in healthcare sector: Causes, consequences and cures. *International Journal of Services and Operations Management*, 1(4), 372-394.

Stecke, K. E., & Aronson, J. E. (1985). Review of operator/machine interference models. *Journal of Production Research*, 23(1), 129-151.

Takagi, H. (1988). Queuing analysis of polling models. *ACM Computing Surveys,* 20(1), 5-28.

Vandaele, N. J., & De Boeck, L. (2003). Advanced resource planning. *Robotics and Computer Integrated Manufacturing,* 19, 211-218.

Vishnevskii, V. M., & Semenova, O. V. (2006). Mathematical methods to study the polling systems. *Automation and Remote Control,* 67(2), 173-220.

Vissers, J. M. H., Bertrand, J. W. M., & De Vries, G. (2001). A framework for production control in health care organizations. *Production Planning & Control*, 12(6), 591-604.

Whitt, W. (1993). Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2), 114-161.

About the authors

Stefan Creemers is professor of operations management at IESEG school of management in France. He is also a visiting professor at HU Brussels. Stefan teaches courses in supply chain management, project management and stochastic modeling and has published award-winning research on these subjects.

Marc Lambrecht is professor of operations management, Research Center for Operations Management, Faculty of Business & Economics, K.U. Leuven, Belgium. He teaches courses in manufacturing systems analysis and inventory management with a focus on stochastic aspects of operations. He has published numerous articles in international journals and he is holder of the Atlas Copco Research Chair in service systems.