Queueing Models in Healthcare

Stefan Creemers
Marc Lambrecht
Nico Vandaele

Abstract - **The healthcare sector is a fast growing segment of GNP in almost every economy. No wonder that we witnessed a tremendous increase in research both medical research to improve medical practice but also research to improve management practices. Patient-flow management is an example. Patient flow represents the ability of the healthcare system to serve patients quickly, reliably and efficiently as they move through stages of care. Unfortunately, patients still experience delays and waiting lists. A queueing model offers an excellent tool to analyze and to improve the performance of healthcare systems. The purpose of this contribution is to discuss differences with the modeling of manufacturing systems and to focus on modeling issues in patient flow. Next, we discuss two specific topics: first, the impact of interrupts and absences on waiting lists and delays, and second, the modeling of batches in healthcare operations.**

1 Introduction

An important feature of healthcare 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. Moreover, timely care is very important so interrupts are common in healthcare processes (the sense of urgency is almost always present). No wonder that healthcare is riddled with delays. 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 more and more 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, ...). If a patient is dismissed before the JLoS is over, the hospital still collects a full reimbursement. On the other hand, if the patient remains in care for a period which exceeds the limit of the JLoS, the hospital has to pay for the extra costs involved. The JLoS of a DRG is determined in function of 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.

Randolph Hall (2006a, 2006b) coined the term patient flow. It represents the ability of the healthcare 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 healthcare. Healthcare 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, imaging machines, laboratories) in any healthcare 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.

Healthcare 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 and 4, we discuss two specific problems, namely, the impact of interrupts and absences on waiting lists and waiting time and the optimization of patient group sizes in medical imaging departments. Section 5 draws some conclusions.

2 Using queueing models to reduce delays in a healthcare system

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

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

The performance measures in healthcare 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, 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. Hereby, 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 and 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 healthcare systems resources are usually dedicated to specific patient types, hospitals have separate units or departments by diagnostic type and bed flexibility is almost non-existing. As a result, capacity pooling is absent. This explains the fact that most queueing models reported in the literature are dealing with parts of the hospital. Queueing models can be used to model hospital wide systems and to evaluate the benefits of greater versus less specialization of care units or other resources (scanners, labs, ...).

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

Time varying demand

Queueing models usually assume time-independent (input) demand rates. Healthcare facilities generally experience different demand over a day, over a week or over 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 healthcare systems we should rely more on time varying arrival rates and time varying server availability and time-dependent waiting times (Green and Soares (2007), Ingolfsson et al. (2002)). Green (2006) proposes a stationary independent period-by-period (SIPP) approach to determine how to vary staffing to meet changing demand.

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 where buffers typically do not consume resources.

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 e.g., 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 re-entry of patients creating additional work on top of the new patients. In most cases, the re-entry is correlated.

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 healthcare 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 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.

Capacity related issues

Hospitals operate within strict business restrictions. Resources are usually very scarce and consequently hospitals operate under high capacity utilization conditions. The so-called heavy traffic conditions are present. Heavy traffic conditions assume that all stations in the network are critically loaded. 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.

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. Natural variability is much more substantial in healthcare 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 non-trivial issue in queueing.

Queueing disciple

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

In what follows, we report on two cases dealing with some of the above mentioned modeling issues.

3 The modeling of interrupts and unplanned absences during healthcare operations

With respect to service outages in healthcare, 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), Chisholm et al. (2000), Chisholm et al. (2001), France et al. (2005) and Gabow et al. (2006) 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 importantly in overtime for staff members. We refer to Easton et al. (2005) for an excellent treatment of this issue.

In this section, we focus on unplanned absences of medical staff and interruptions during service operations. Unplanned absences and interruptions during service activities have a major impact on flow times. Doctors and medical staff face various obligations which they have to attend to (making morning rounds, answering phones, patient check-ups, daily management, ...). In addition doctors often combine a hospital job and private consultation. These phenomena may cause a variable arrival pattern at the hospital (Liu et al., 1998) and may lead to interruptions during the treatment process (Chisholm et al., 2000; Chisholm et al., 2001; Easton et al., 2005; Gabow et al., 2006). It is clear that hospital environments are characterized by substantial amounts of variability. As is argued in the literature (Hopp and Spearman, 2000), variability induces waiting times. While in service industries variability cannot be countered by means of inventory in the traditional sense, patients will have to wait until capacity becomes available (Vissers, Bertrand and De Vries, 2001; Harper, 2002; Vandaele and De Boeck, 2003; Sethuraman and Tirupati, 2005). Besides 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. Over the past decades, 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, imagine a doctor's office that 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

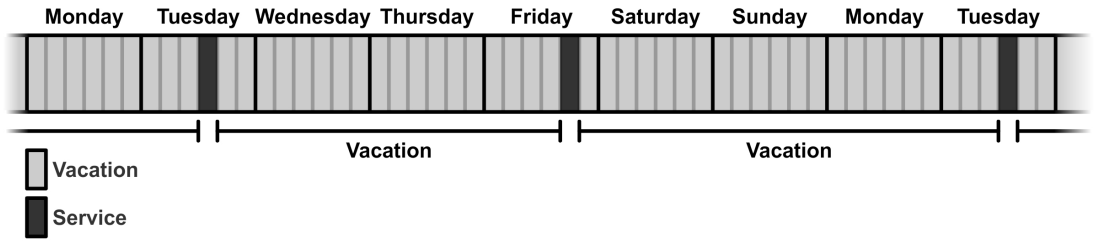

Figure 1: Illustration of a vacation model

by Hopp and Spearman (2000). In their work, Hopp et al. (2000) propose a transformation of the service process times to account for service outages. The results of Hopp et al. (2000) are widely accepted in the literature (see for instance Lambrecht, Ivens and Vandaele (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. Next, we provide the means to model them.

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.

We distinguish between preemptive and nonpreemptive outages. Preemptive and nonpreemptive outages will impact 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_T^2 respectively (i.e., absence times are allowed to follow a general distribution). Furthermore, we assume an average number of patients (represented by n) to 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 which builds on the tradition set by Hopp et al. (2000). They are characterized by a Mean Time To Interrupt (τ_i) and a Mean

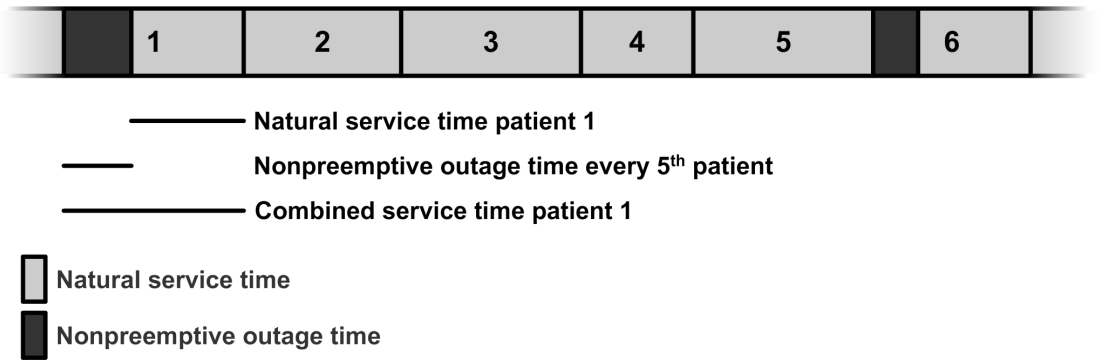

Figure 2: Illustration of the combined service time

Time To Resolve (τ_r). The model presented in Hopp et al. (2000) presumes interrupts to 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, ...

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 occurence of nonpreemptive and preemptive outages.

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, setups, rework, maintenance, ... are all extensions that are able to capitalize on the technique discussed in this section). Applied to our setting, we have 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_T^2 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:

$$f_c(x+y) = f(x)f_T(y) \tag{1}$$

Where $f(x)$ is the probability density function of the natural service times with mean $\overline{X}$ and variance s_X^2. 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:

$$\begin{aligned} \overline{X}_s &= \left[\left(\frac{n-1}{n}\right)\int f(x)x\mathrm{d}x\right] + \left[\frac{1}{n}\iint f(x)f_T(y)(x+y)\mathrm{d}y\mathrm{d}x\right] \\ &= \overline{X} + \frac{T}{n} \end{aligned} \tag{2}$$

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

$$\begin{aligned} s_s^2 &= \left[\left(\frac{n-1}{n}\right)\int f(x)\left(x-\overline{X}_s\right)^2\mathrm{d}x\right] + \left[\frac{1}{n}\iint f(x)f_T(y)\left(x+y-\overline{X}_s\right)^2\mathrm{d}y\mathrm{d}x\right] \\ &= s_X^2 + \frac{s_T^2}{n} + T^2\left(\frac{n-1}{n^2}\right) \end{aligned} \tag{3}$$

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

Preemptive outages

We refer to service interruptions as preemptive outages. Doctors being called away on emergencies, answering phone calls, ... 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 while 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 being 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 to occur only during actual service time. As such, preemptive outages do not take place during the resolve times induced by previous outages. Remark 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 be interrupted when already engaged in resolving a previous interrupt). While this latter instance can be seen as an extension of the former, we will first discuss outages occurring exclusively during actual service time. Define $\tau_{r_0}(j)$ 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_r^2. In addition, resolve times of different outages are assumed to be i.i.d. random variables. 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:

$$\overline{X}_i = \overline{X} + \sum_{j=1}^{J_0}\tau_{r_0}(j) \tag{4}$$

As such, the average service time $\overline{X}_i$ incorporates both the natural service time $\overline{X}$ 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 $\tau_{r_0}(j)$) in which the number of random variables (the number of interrupts J_0), is a random variable itself. Assume that J_0 and $\tau_{r_0}(j)$ are i.i.d. variables for all $j \in \mathbb{N}_0$. In addition, assume the mean as well as the variance of to be equal for all $j \in \mathbb{N}_0$. Therefore, the mean and variance of the sum of J_0 random variables can be expressed as (Dudewicz and Mishra, 1988):

$$E[S_0] = E[J_0]\, E[\tau_{r_0}(j)] \tag{5}$$

$$s_{S_0}^2 = E[J_0]\, s_r^2 + E[\tau_{r_0}(j)]^2\, s_{J_0}^2 \tag{6}$$

Where S_0 is the random variable representing the sum of J_0 resolve times $\tau_{r_0}(j)$. In other words, we have that:

$$S_0 = \sum_{j=1}^{J_0} \tau_{r_0}(j) \tag{7}$$

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

$$E[S_0] = \overline{X}\left(\frac{\tau_r}{\tau_i}\right) \tag{8}$$

$$s_{S_0}^2 = \overline{X}\left(\frac{s_r^2 + \tau_r^2}{\tau_i}\right) \tag{9}$$

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

$$\overline{X}_i = \overline{X}\left(\frac{\tau_i + \tau_r}{\tau_i}\right) \tag{10}$$

This corresponds to the expression presented in Hopp et al. (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 time. We start with the expression of the second moment:

$$E\left[X_i^2\right] = \left[\left(s_X^2 + \overline{X}^2\right)\left(1 + \frac{\tau_r}{\tau_i}\right)^2\right] + s_{S_0}^2 \tag{11}$$

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

$$s_i^2 = \left[s_X^2\left(1 + \frac{\tau_r}{\tau_i}\right)^2\right] + s_{S_0}^2 \tag{12}$$

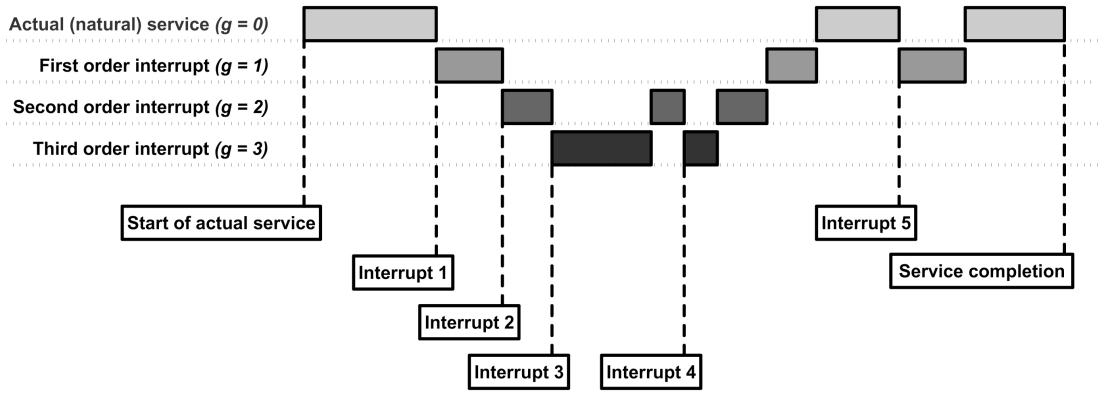

Figure 3: Multilevel interruptions during the service process of a patient

This expression once more matches the formula derived in Hopp et al. (2000). The above 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, ... In 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:

$$S_g = \sum_{j=1}^{J_g} \tau_{r_g}(j) \tag{13}$$

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:

$$E\left[J_g\right] = E\left[J_g^2\right] - E\left[J_g\right]^2 = \overline{X}\frac{1}{\tau_i}\left(\frac{\tau_r}{\tau_i}\right)^g \tag{14}$$

One can infer that:

$$E\left[S_g\right] = \overline{X}\left(\frac{\tau_r}{\tau_i}\right)\left(\frac{\tau_r}{\tau_i}\right)^g \tag{15}$$

$$s_{S_g}^2 = \overline{X}\left(\frac{s_r^2+\tau_r^2}{\tau_i}\right)\tau_r\left(\frac{\tau_r}{\tau i}\right)^g \tag{16}$$

Using the same reasoning as applied previously, one can express the mean service time

including the effect of all order interrupts as follows:

$$\overline{X}_i = \overline{X}\left(\frac{\tau_i}{\tau_i - \tau_r}\right) \tag{17}$$

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

$$E\left[X_i^2\right] = \left\{\left(s_X^2 + \overline{X}^2\right)\left[1 + \frac{2\tau_r}{\tau_i - \tau_r} + \left(\frac{\tau_r}{\tau_i - \tau_r}\right)^2\right]\right\} + \overline{X}\left(\frac{s_r^2 + \tau_r^2}{\tau_i - \tau_r}\right) \tag{18}$$

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

$$s_i^2 = \frac{s_X^2\tau_i^2 + \overline{X}\left(\tau_i - \tau_r\right)\left(s_r^2 + \tau_r^2\right)}{\left(\tau_i - \tau_r\right)^2} \tag{19}$$

Combining preemptive and nonpreemptive outages

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

$$\begin{aligned} \overline{X}_{T_i} &= \left[\left(\frac{n-1}{n}\right)\int f_i(x)x\mathrm{d}x\right] + \left[\frac{1}{n}\iint f_i(x)f_T(y)\left(x+y\right)\mathrm{d}y\mathrm{d}x\right] \\ &= \overline{X}_i + \frac{T}{n} \end{aligned} \tag{20}$$

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_i^2 respectively. We refer to $\overline{X}_{T_i}$ 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:

$$\begin{aligned} s_{T_i}^2 &= \left[\left(\frac{n-1}{n}\right)\int f_i(x)\left(x - \overline{X}_{T_i}\right)^2\mathrm{d}x\right] + \left[\frac{1}{n}\iint f_i(x)f_T(y)\left(x+y-\overline{X}_{T_i}\right)^2\mathrm{d}y\mathrm{d}x\right] \\ &= s_i^2 + \frac{s_T^2}{n} + T^2\left(\frac{n-1}{n^2}\right) \end{aligned} \tag{21}$$

4 The batching decision for the nuclear magnetic resonance scanner

The Nuclear Magnetic Resonance scanner (named NMR hereafter) is one of these very expensive high tech resources, used for medical imaging. In this section we will analyze how

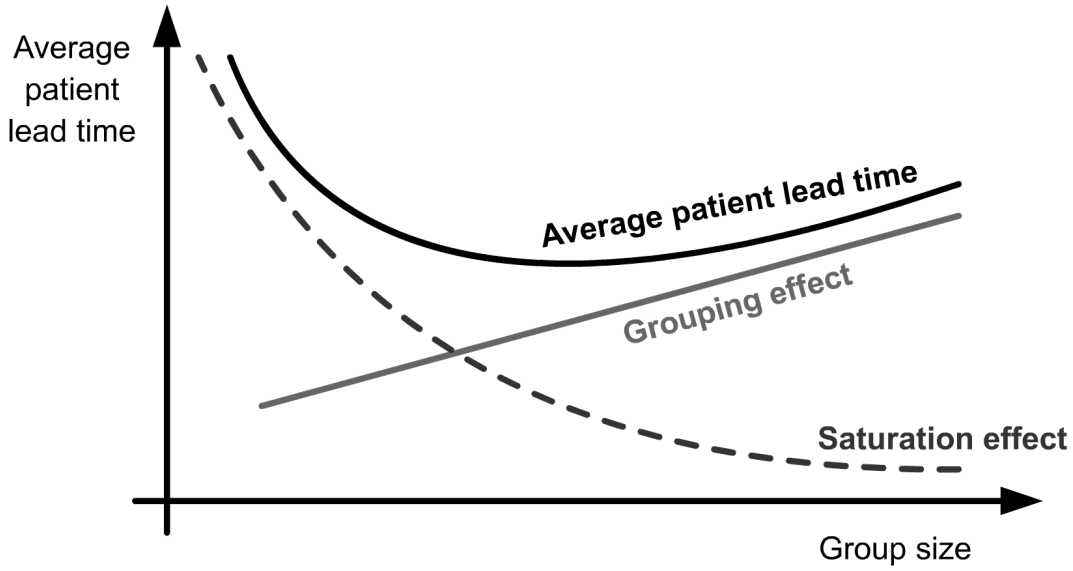

Figure 4: Convex relationship between average patient lead time and group size

the use of this resource can be managed in a more efficient way, by controlling the group sizes of different patient types in such a way that the weighted average patient lead time is minimized. As will be shown, this does not only improve patient comfort (in terms of total time spent in the system), but also the responsiveness of the system (in terms of availability in case of emergency calls).

The NMR technology is based on three subsequent steps: nuclear magnetization, resonance and relaxation. The NMR equipment consists of a strong electromagnet and a set of measurement antennas. The electromagnet is cylindrical, allowing the accommodation of the appropriate part of the human body. At the central point of the cylinder, the magnetic field is the most homogeneous.

The antennas are specifically developed in function of the part of the human body to be measured. As a consequence, each time another type of image is needed, the system has to switch to the appropriate antenna or the antenna has to be set up differently. In either case, a combination of both hardware and software setups has to be performed (Truyen, 1996).

Because of these setups, the grouping of patients is a common practice in medical imaging departments. Although the need for grouping is quite obvious, determining the optimal size of the groups is not self-evident. On the one hand there is a need to use the NMR equipment efficiently, which means that the number of setups must be kept low and group sizes should be large. On the other hand, patients cannot wait too long before being processed, as patient lead time has become an important element of competition between hospitals. This argument drives for small group sizes.

In order to explain this inherent conflict, we analyze the relationship between the average patient lead time (defined as the time between a patient's call arrival at the NMR department and the moment at which he leaves and the group size. This relationship is represented in Figure 4.

The two conflicting effects mentioned above are referred to as the grouping effect and the saturation effect (Karmarkar, 1987). The grouping effect (straight line on Figure 4) shows increasing average patient lead times when groups are larger. The larger the group, the longer the group lead time, and subsequently the longer the patients will remain in the system. The saturation effect (dotted line on Figure 4) shows increasing lead times when groups are smaller. Smaller groups require more frequent setups, and as a consequence, the utilization of the NMR equipment increases. This causes congestion, resulting in longer average lead times.

The combination of both effects results in a convex relationship (Lambrecht, Chen and Vandaele, 1996), which implies that there is an optimal group size minimizing average patient lead time. More details can be found in Vandaele, 1996. In the following paragraph, we develop the mathematical model we will use to address the group size decision problem. The objective is to determine the group size that minimizes the average patient lead time. Clearly, this is not only desirable from a patient's point of view (since it improves patient comfort), but also enhances the responsiveness of the system.

Model description

In our approach, the NMR is regarded as a multi-product, single server open queueing model. As such, it is an extension of the single product, single server model described in Kraemer and Lagenbach-Belz (1976). If we view a request for a particular image (part of the human body) as an arrival of that image type, we can define the parameters of the model as follows:

k = image type index, $k \in \{1, \ldots, K\}$
Y_k = average interarrival time of image type k
$s^2_{Y_k}$ = interarrival time variance of image type k
$c^2_{Y_k}$ = squared coefficient of variation of the interarrival time of image type k
λ_k = average arrival rate of image type k
$\overline{T}_k$ = average setup time for image type k
$s^2_{T_k}$ = setup time variance for image type k
$c^2_{T_k}$ = squared coefficient of variation of the setup time of image type k
$\overline{X}_k$ = average unit processing time of image type k
$s^2_{X_k}$ = unit processing time variance for image type k
$c^2_{X_k}$ = squared coefficient of the unit processing time of image type k
μ_k = unit processing rate of image type k

In order to construct the model, the parameters of the individual arrival and service processes are aggregated into a single aggregate arrival and service process. Simultaneously, we integrate the grouping aspect of the service process into our model.

Objective function and optimization

Given the average group arrival rate λ_k and average group processing rate μ_k, the expected lead time for each image type k can be obtained as follows:

$$E[W_k] = \frac{Q_k - 1}{2\lambda_k} + E[W_q] + \overline{T}_k + \frac{Q_k + 1}{2\mu_k} \tag{22}$$

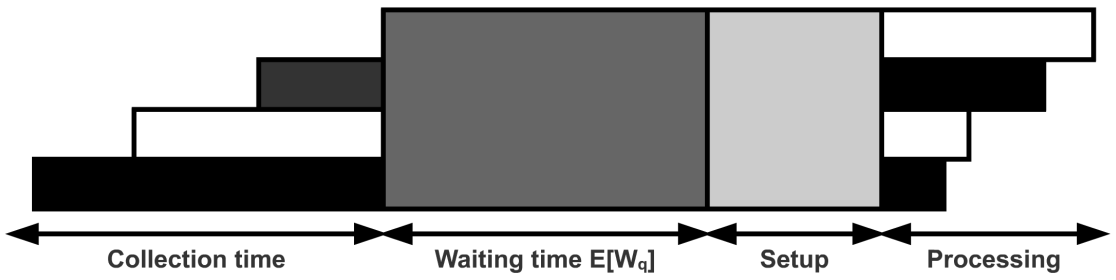

Figure 5: Visualization of the different phases of the group lead time

This lead time clearly consists of four building blocks. The first term corresponds to the average time a patient of image type k will have to wait until a group of size Q_k has been formed (collection time). The term $E[W_q]$ represents the average time that patients spend waiting in queue in front of the scanner until it becomes idle (waiting time). It is approximated by the Kraemer-Lagenbach-Belz formula (Kraemer and Lagenbach-Belz, 1976), and is independent of the image type. The last two terms correspond to the average time a patient of image type k spends in setup and processing. These different phases are visualized in Figure 5 for a group size of four patients. The model assumes a FIFO discipline, which is accepted to be fair among patients in a waiting room.

As can be seen from the figure, the collection time is different for each patient of the group: the first patient has to wait until the other three have arrived, while after arrival of the last patient the entire group is immediately transferred to the second stage namely the queue in front of the scanner. The time spent in queue and setup is the same for each patient of the group. The time spent in processing is again different: given the FIFO discipline and the fact that the scanner can deal with only one patient at a time, the first patient will be processed immediately while the others will have to wait until the preceding patient has left the system.

The aggregate average lead time $E[W]$ is calculated as a weighted average of the lead times of all individual image types:

$$E[W] = E[W_q] + \sum_{k=1}^{K} \frac{\lambda_k}{\sum_{k=1}^{K} \lambda_k} \left[\frac{Q_k - 1}{2\lambda_k} + \overline{T}_k + \frac{Q_k - 1}{2\mu_k} \right] \tag{23}$$

The weights are chosen to reflect the relative importance of each image type for the system. Here, the importance is measured by means of the volume of the image type. Alternatively, other weighting factors could be applied.

At this point the model is completed and we can formally state our optimization problem:

$$\min E[W] = E[W_q] + \sum_{k=1}^{K} \frac{\lambda_k}{\sum_{k=1}^{K} \lambda_k} \left[\frac{Q_k - 1}{2\lambda_k} + \overline{T}_k + \frac{Q_k - 1}{2\mu_k} \right] \tag{24}$$

k	Name	Abbreviation
1	Skull/Foot/Ankle	SFA
2	Lumbal spine	LS
3	Cervical spine	CS
4	Shoulder/Hip	SH
5	Knee/Wrist/Elbow	KWE
6	Rest (neck, breast, ...)	REST

Table 1: Overview of the six image types of the NMR scanner

s.t.

$$\rho_e < 1 \tag{25}$$

$$O_k \geq 1 \tag{26}$$

Where ρ_e equals the effective traffic intensity. It is a measure of the load of the scanner, and can be derived from the traditional traffic intensity ρ by adding the effect of the setup times:

$$\rho_e = \frac{l}{k} = \sum_{k=1}^{K} l_k \left[\overline{T}_k + Q_k \overline{X}_k\right] = \sum_{k=1}^{K} l_k \overline{T}_k + \rho \tag{27}$$

Where ρ is defined as:

$$\rho = \sum_{k=1}^{K} \rho_k = \sum_{k=1}^{K} \frac{\lambda_k}{\mu_k} \tag{28}$$

In order to preserve system stability, the effective traffic intensity ρ_e should be strictly smaller than unity. It is clear that ρ_e is dependent on the group sizes Q_k. For large group sizes, l_k tends to zero and ρ_e approaches the traditional traffic intensity ρ. For small group sizes, ρ_e increases due to the impact of the setup times.

This non-linear constrained optimization problem is solved using a dedicated optimization algorithm, yielding the optimal group size Q_k^* for each image type k. The expected individual lead time for a patient of image type k can then be obtained as follows:

$$E\left[W_k\right]_{\text{OPT}} = \frac{Q_k^* - 1}{2\lambda_k} + E\left[W_q\right](Q_k^*) + \overline{T}_k + \frac{Q_k^* + 1}{2\mu_k} \tag{29}$$

We shortly summarize an example of this approach, based on a realistic data set (see also Vandaele, Van Nieuwenhuyse and Cuypers, 2003). The NMR scanner produces ten types of images, only five of which occur frequently. In Table 1 those five image types are described (Image type 1 to 5). The infrequent image types are summarized in Image type 6.

In this example there is no distinction between internal patients (coming from another hospital department) and external patients (coming from outside). The time needed to carry out the measurements is dependent on the image type (ranging from 12 seconds to 7

k	$\overline{Y}_k$ (sec)	$\overline{Y}_k$ (min)	$s^2_{Y_k}$ (sec^2)	$c^2_{Y_k}$	λ_k (per sec)
1	14400	240	288376200	1.391	0.0000694
2	6120	102	25416000	0.679	0.0001634
3	7971	133	144803314	2.279	0.0001254
4	27390	457	353248200	0.471	0.0000365
5	5375	90	16207500	0.561	0.0001860
6	17850	298	477405000	1.498	0.0000560

Table 2: Summary of the arrival data for the different image types

k	$\overline{T}_k$ (sec)	$s^2_{T_k}$ (sec^2)	$c^2_{T_k}$
1	46.714	338.905	0.155
2	89.375	1745.982	0.219
3	77.857	882.143	0.146
4	80.000	3200.000	0.500
5	91.000	2318.667	0.280
6	110.000	7725.000	0.638

Table 3: Summary of the setup time data for the different image types

minutes) and the number of measurements (ranging from 3 to 8). A rule of thumb is that the smaller the part of the human body, the more measurements have to be performed. The patient also influences the number of measurements: nervousness or movements can cause the measurement to be redone, increasing processing variability.

The NMR scanner was observed during several weeks in order to collect the necessary data. The arrival data (for individual patients) are summarized in Table 2. As can be seen, there are significant differences in the frequencies of occurrence for the individual image types. In addition, the image types seem to exhibit a significant different variability (squared coefficients of variation).

The observed setup time and processing time characteristics are given in Table 3 and Table 4.

For all image types, the average setup and processing times are in the same order of magnitude. The processing times exhibit significantly less variability than the setup times. In order to take into account the different outages of the system, such as lunches, meetings, working schedules and interruptions, we have to rely on effective processing and setup times, based on the availability concept. Without going into the details, these data are given in Table 5.

At this point, we have all the necessary data to run the optimization model. The results are discussed in the next section.

k	$\overline{X}_k$ (sec)	$\overline{X}_k$ (min)	$s^2_{X_k}$ (sec^2)	$c^2_{X_k}$
1	1525	25	244445.450	0.1050
2	1164	19	113157.143	0.0836
3	1103	18	64740.000	0.0533
4	1428	24	34920.000	0.0170
5	948	16	53840.390	0.0600
6	1640	27	37200.000	0.0138

Table 4: Summary of the processing time data for the different image types

k	Effective average setup time (sec)	Effective setup time variance (sec^2)	Effective average processing time (sec)	Effective processing time variance (sec^2)
1	56.910	503.000	1857.868	362803.935
2	108.883	2591.372	1417.548	167946.905
3	94.851	1309.2769	1343.147	96086.578
4	97.462	4749.414	1739.695	51827.978
5	110.863	3441.347	1154.424	79909.466
6	134.010	11465.382	1997.970	55211.935

Table 5: Effective setup and processing time data

k	Planned group sizes		Actual group sizes		Optimal group sizes	
	(+)	(−)	(+)	(−)	(+)	(−)
Overall	8	7	3	2	NA	NA
1	8	7	2	1	2	1
2	6	6	4	3	2	1
3	7	7	2	2	3	2
4	5	5	3	2	2	1
5	13	12	5	4	2	1
6	5	5	1	1	2	1

Table 6: Optimal group sizes versus planned and actual group sizes

Optimization and results

The resulting optimal group sizes (rounded above (+) and rounded below (−)) are shown in Table 6, along with the planned group sizes and the actual group sizes as currently used in practice. The planned group sizes are the group sizes used by the administration to list incoming calls/arrivals whereas the actual group sizes represent the registered practice. In this way, the planned and the actual group sizes represent the group sizes with and without the unplanned emergency calls. Given the convex relationship from Figure 4, rounding to the nearest larger integer is a conservative approach. On the contrary, rounding the nearest lower integer is opportunistic but can in certain situations lead to infeasibilities (due to the effect of the increased time spent on setups, which may lead to an adapted traffic intensity equal to or larger than unity).

The table shows a significant difference between the planned group sizes and the actual group sizes. The reasons for this are diverse: emergencies, patient related causes (e.g., late arrivals, cancellations), wrong diagnosis by the corresponding doctor, etc. As the resulting actual group sizes are always smaller than the planned group sizes, the actual number of setups performed is larger than planned. Furthermore, Table 6 shows that the optimal group sizes (as calculated by the model) confirm the actual batch sizes, and even indicate opportunities for further group size reduction for image types 3, 4 and 5.

Table 7 gives an overview of the respective average patient lead times, as well as the percentage improvement in average patient lead time that could be obtained by introducing the optimal group sizes (as calculated by the model) instead of the actual group sizes. We can conclude that group sizes of 2 would be optimal, except for type 3 which needs a group size of 3 (these optimal group sizes are all rounded to the larger integer, as rounding to the smaller integer leads to infeasibilities). As can be seen, the percentage lead time improvements range between 5 % for type 6 patients up to 48 % for type 5 patients. As such, it is clear that using these optimal group sizes would enhance patient comfort and in the meantime improve the availability of the NMR in case of emergency operations. Since the group sizes used in practice are mainly a management decision, it can be concluded that the model provides hospital management with powerful and yet easy to implement guidelines for efficiency improvement.

k	Planned group size	Planned LT (min)	Actual group size	Actual LT (min)	Optimal group size	Optimal LT (min)	Lead time improvement
1	8	1671	2	753	2	539	29%
2	6	981	4	820	2	456	44%
3	7	1139	2	683	3	558	19%
4	5	1642	3	1115	2	644	42%
5	13	1371	5	847	2	441	48%
6	5	1346	1	606	2	574	5%

Table 7: Average patient lead times in seconds (minutes) for the planned, the actual and the optimal group sizes

5 Conclusion

Queueing models are very useful to quantify the relationship between capacity utilization, waiting time and patient (customer) service. Increasing the capacity utilization, while financially attractive, may require unacceptably high patient delays and poor customer service. The presence of high degrees of variability (outages, absences, interruptions, ...) negatively impacts the performance of the healthcare system. In this paper, we show that this capacity and variability analysis in a healthcare environment results in queueing models that are different from queueing models in an industrial setting. This is illustrated by elaborating on two case studies, one case study is related to interrupts and absences and a second case study is dealing with a batching decision for a Nuclear Magnetic Resonance Scanner.

References

Babes, M. and Sarma, G.V., 1991, Out-Patient Queues at the Ibn-Rochd Health Center, The Journal of the Operational Research Society, 42, 10. 845-855.

Brandeau, M., Sainfort, F. and Pierskalla, W., 2004, Operations Research and Health Care, (Kluwer Academic Publishers).

Chisholm, C.D., Collison, E.K., Nelson, D.R. and 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. and 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. and Lambrecht, M.R., 2007, Modeling a Healthcare System as a Queueing Network: The Case of a Belgian Hospital, (Research Report 0710, Department of Decision Sciences & Information Management, Research Center for Operations Management, KU

Leuven).

Doshi, B.T., 1986, Queueing Systems with Vacations - A Survey, Queueing Systems: Theory and Applications, 1, 1, 29-66.

Easton, F.F. and Goodale, J.C., 2005, Schedule Recovery: Unplanned Absences in Service Operations, Decision Sciences, 36, 3, 459-488.

Federgruen, A. and Green, L., 1986, Queueing Systems with Service Interruptions, Operations Research, 34, 5, 752-768.

France, J.D., Levin, S., Hemphill, R., Chen, K., Rickard, D., Makowski, R., Jones, I. and Aronsky, D., 2005, Emergency Physicians' Behaviors and Workload in the Presence of an Electronic Whiteboard, International Journal of Medical Informatics, 74, 827-837.

Gabow, P.A., Karkhanis, A., Knight, A., Dixon, P., Eisert, S. and Albert, R.K., 2006, Observations of Residents' Work Activities for 24 Consecutive Hours: Implications for Workflow Redesign, Academic Medicine, 81, 8, 766-775.

Green, L., 2006, Queueing Analysis in Healthcare, in Hall, R. ed.,Patient Flow: Reducing Delay in Healthcare Delivery (Springer Science), 281-307.

Green, L. and 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 (Springer Science).

Hall, R., 2006b, 2006, Patient Flow: The New Queueing Theory for healthcare, OR/MS Today, 23, 3, 36-40.

Haque, L. and Armstrong, M.J., 2007, A Survey of the Machine Interference Problem, European Journal of Operational Research, 179, 469-482.

Harper, P.R., 2002, A Framework for Operational Modelling of Hospital Resources, Health Care Management Science, 5, 165-173.

Hopp, W. and Spearman, M., 2000, Factory Physics, Foundations of Manufacturing Management, (Irwin/McGraw-Hill, New York).

Ingolfsson, A., Haque, A. and Umnikov, A., 2002, Accounting for Time-Varying Queueing Effects in Workforce Scheduling, European Journal of Operational Research, 139, 585-597.

Karmarkar, U., 1987, Lot Sizes, Lead Times and In-Process Inventories, Management Science, 33, 3, 409 - 423.

Kraemer, W. and Lagenbach-Belz, M., 1976, Approximate Formulae for the Delay in the Queueing System GI/GI/1, (Congressbook, Eighth International Teletraffic Congress, Melbourne), 235-1/8.

Lambrecht, M.R., Chen, S. and Vandaele, N.J., 1996, A Lot Sizing Model with Queueing Delays. The Issue of Safety Time, European Journal of Operational Research, 89, 2, 269-276.

Lambrecht, M.R., Ivens, P.L. and Vandaele, N.J., 1998, Aclips: A Capacity and Lead Time Integrated Procedure for Scheduling, Management Science, 44, 11, 1548-1561.

Liu, L. and Liu, X., 1998, Block Appointment Systems for Outpatient Clinics with Multiple Doctors, 49, 12, 1254-1259.

Sethuraman, K. and 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. and 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.

Truyen, L., 1996, Magnetic Resonance Imaging Studies in Multiplesclerosis, (PhD thesis, Medical Department, Universitaire Instellingen Antwerpen, Antwerp).

Vandaele, N., 1996, The Impact of Lot Sizing on Queueing Delays: Multi-Product, Multi-Machine Models, (PhD Thesis, Department of Applied Economics, KU Leuven, Leuven).

Vandaele, N., Van Nieuwenhuyse, I. and Cuypers, S., 2003, Optimal Grouping for a Nuclear Magnetic Resonance Scanner by Means of an Open Queueing Model, European Journal Of Operational Research, 151, 181-192.

Vandaele, N.J. and De Boeck, L., 2003, Advanced Resource Planning, Robotics and Computer Integrated Manufacturing, 19, 211-218.

Vishnevskii, V.M. and 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. and De Vries, G., 2001, A Framework for Production Control in Health Care Organizations, Production Planning & Control, 12, 6, 591-604.