KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Healthcare queueing models

Stefan Creemers and Marc Lambrecht

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0804

Healthcare queueing models

Stefan Creemers[1] and Marc Lambrecht[1]

[1] Faculty of Business and Economics

Department of Decision Sciences and Information Management

Catholic University Leuven, Belgium

e-mail: firstname.lastname@econ.kuleuven.be

Abstract

Healthcare systems differ intrinsically from manufacturing systems. As such, they require a distinct modeling approach. In this article, we show how to construct a queueing model of a general class of healthcare systems. We develop new expressions to assess the impact of service outages and use the resulting model to approximate patient flow times and to evaluate a number of practical applications. We illustrate the devastating impact of service interruptions on patient flow times and show the potential gains obtained by pooling hospital resources. In addition, we present an optimization model to determine the optimal number of patients to be treated during a service session.

Keywords:

- Operations Research [MeSH: L01.906.575]
- Efficiency, Organizational [MeSH: N04.452.227]
- Health Care Evaluation Mechanisms [MeSH: N05.715.360]
- Decision Support Systems, Management [MeSH: N04.452.515.135]
- Time Management [MeSH: N04.452.932]
- Queueing Theory

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. DRG's (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 [1]. Delays also create a "hidden" hospital in analogy with the hidden company. In other words, such a hospital creates wasteful overhead.

Hall et al. [2] 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.

In this article we try to address some of the issues mentioned above. The contribution of this article is threefold: (1) we provide the tools to model a complex hospital system; (2) we develop new expressions to assess the impact of service outages; (3) we present a number of practical applications. More specifically, we demonstrate the impact on system performance resulting from the reduction of service outages and illustrate the beneficial effects of pooling. Moreover, we develop an optimization model that enables us to determine the optimal number of patients to be treated during a service session (e.g. a consultation time block). The remainder of this article is structured as follows: in section 2 we discuss features that make the modeling of a hospital system more difficult than the modeling of a typical industrial manufacturing process. In section 3 we model a hospital queueing system. In section 4 we provide a numerical example and we discuss a number of practical applications. Section 5 concludes.

2 Problem description

Queueing models have been applied in numerous industrial settings and service industries. The number of applications in healthcare, however, is 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 [3]. The second modeling issue has to do with 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, 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.

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.

Service sessions 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 "service sessions" 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 sessions. Vacation models observe the queueing behavior of such systems in

which servers are available during certain time intervals 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, inaccurate results have a large impact on resulting performance measures.

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.

3 A hospital queueing system

The features discussed in the previous section considerably complicate the modeling exercise. In order to demonstrate how to implement the features in a queueing model, we use an example hospital queueing system. The example concerns a typical hos-

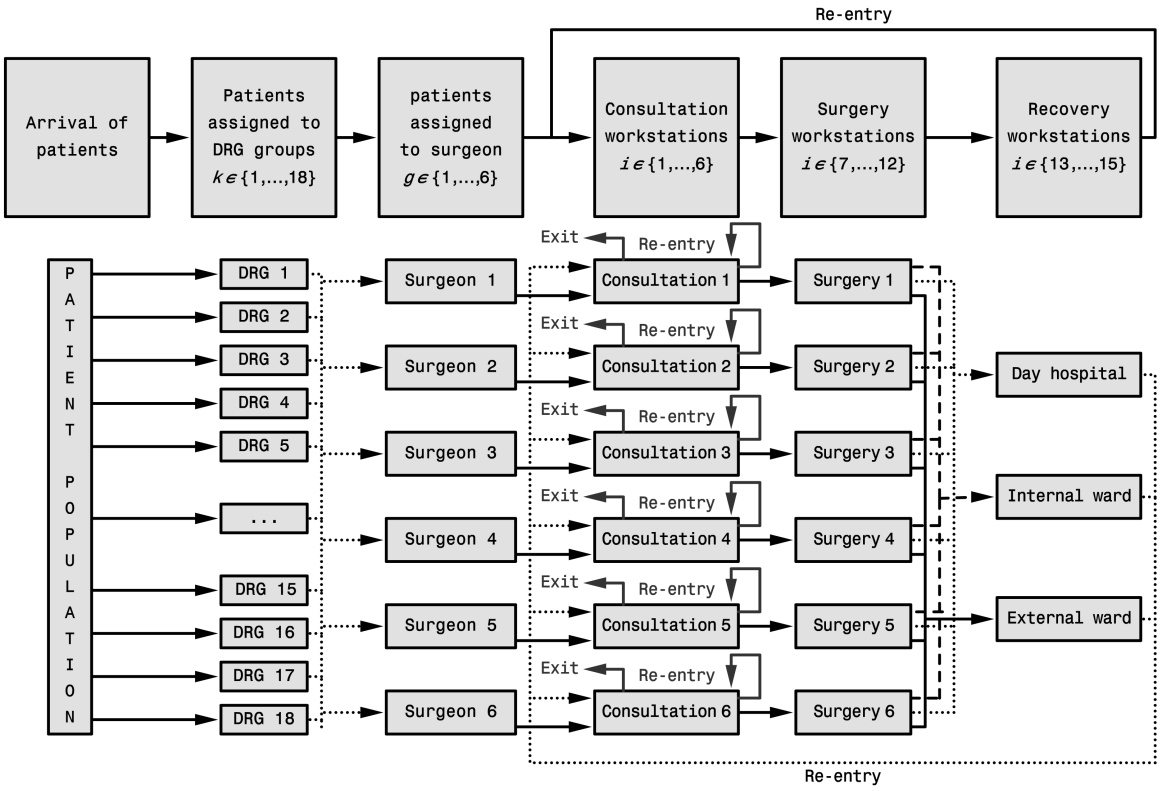

Figure 1: Capacity structure of the hospital department

pital department involving consultation, surgery and recovery. The example we use throughout this paper is inspired by a real life case of the orthopedic department of the Middelheim hospital (Antwerp, Belgium) [4]. We omit in this paper all practical data collection details of the case. We now and then provide numerical data to give the reader an idea of the problem dimension. In our example, the department employs six surgeons. Each of the surgeons is assigned a certain number of patients and no patient crossover between surgeons is assumed to take place. The base case deals in other words with the non-pooled capacity. Recovery occurs in an internal ward, in an external ward or in the day hospital (depending on the disorder the patient is suffering from). In each of the wards 25 beds are reserved for patients of the hospital department under study. The capacity structure of the department is illustrated in Figure 1. Notwithstanding the fact that every patient is unique, we impose some general assumptions regarding

the treatment process of a patient visiting the department. More specifically, we assume that every patient starts the treatment process with one or more consultations. Next, surgery is performed and a number of follow-up consultations is initiated. Finally the treatment process of a patient finishes and the patient leaves the hospital system. We assume that only elective surgery takes place and that the consultation process is appointment-based. Remark that it is possible to specify other patient routings (e.g. patients who refuse surgery, patients that do not longer need recovery, ...). In this example however, we make use of a simple patient routing structure in order to preserve the transparency of the model.

With respect to the performance measures, we are interested in the total flow time of a patient at a workstation (i.e. consultation, surgery or recovery). We define the flow time as the total waiting time plus the processing time. With respect to the waiting time of a patient, a distinction is made between the internal waiting time and the external waiting time (for instance refer to Vissers et al. [5] and Hall et al. [1]). More specifically, the internal waiting time is the time spent inside the hospital prior to receiving service (at any of the workstations). The external waiting time is the time between the making of an appointment and the arrival of a patient at the hospital. The external waiting time can also be related to the "waiting list" phenomenon. As such, the total flow time of a patient consists of: (1) the external waiting time; (2) the internal waiting time; (3) the processing time. In the remainder of this text we will use $E[W]$ to denote the expected total flow time of a patient.

The data collection may be described in the following way (see also Figure 1). We start with a patient population (in our case we collected data on the consultation, surgery and recovery process of 3,300 patients) and divide it into groups of similar DRG's. We construct 18 DRG groups and use index k, $k \in \{1, 2, \ldots, K\}$ for further identification (refer to Roth et al. [6] and van Merode et al. [7] for a detailed treatment

on patient classification methodology). Next, the patients are assigned an individual surgeon (identified using index $g,\ g \in \{1, 2, \ldots, G\}$). Surgeons as well as recovery wards may be considered as hospital resources. We use index $i,\ i \in \{1, 2, \ldots, I\}$ to identify these resources. The surgeons perform both consultation ($i \in \{1, 2, \ldots, 6\}$) as well as surgery ($i \in \{7, 8, \ldots, 12\}$) tasks. Recovery takes place at the day hospital ($i = 13$), at the internal ward ($i = 14$) or at the external ward ($i = 15$).

In what follows we develop the queueing model. First we provide the mathematical derivations required to obtain the arrival- and natural process times. Next, we adapt the model to include the effects of service outages, the availability of workstations and the characteristics of the aggregate arrival process.

3.1 Modeling arrival rate and natural service times

The queueing model of the hospital department may be presented as a network of 12 $G/G/1$ workstations (six surgeons performing both consultation and surgery) and 3 $G/G/m$ workstations (the recovery wards). The network is an open re-entry network with stochastic routings and is modeled using the principles of the parametric decomposition approach that was pioneered by Jackson [8]. The parametric decomposition approach has further been refined by authors such as Kingman [9], Shanthikumar et al. [10], Bitran et al. [11], Whitt [12], Lambrecht et al. [13] and Vandaele et al. [14]. While other approaches are available (e.g. Brownian motion queueing models), a previous study has shown that the parametric decomposition approach works best when modeling complex hospital systems [4].

The queue discipline adhered at each of the stations is FCFS. Any variation in the arrival of patients (e.g. the early, late, unannounced or not showing up of patients) is presumed to be absorbed in the variance of the arrival process. The model assumes infinite buffers to exist in front of every queue. Realizing that the buffers in front of

the consultation and surgery workstation correspond to their respective waiting lists, it would be incorrect to restrain them in size. In real life, if patients contact the hospital to make an appointment for a consultation or a surgery, they will be issued an appointment date no matter how far ahead in time this date might be (i.e. we assume patients not to display any balking- or reneging-behavior when arriving or abiding at the queue). Hence buffer capacities are virtually unlimited. With respect to the recovery wards, one might argue that queue capacity is in fact limited. However, there are several reasons that are able to question this assertion. Next to rendering the model highly intractable, finite buffers do not necessarily correspond to reality since shortages of bed capacity at the wards are solved at the local level and in general do not prolong the sojourn time of a patient (this of course presumes the presence of unoccupied beds somewhere in the hospital). Therefore we will assume infinite buffers at all stages of the treatment process. Considering the multiclass re-entry environment of the queueing network, aggregation of the arrival and service process is required in order to perform a decomposition-based queueing analysis.

More formally, let i ($i \in \{1, \ldots, I\}$) denote the workstation in the network, let k ($k \in \{1, \ldots, K\}$) denote the DRG group a patient belongs to and let g ($g \in \{1, \ldots, G\}$) denote the surgeon a patient is assigned to. As such, we have KG classes of patients visiting a set of I workstations. Let the pair (k, g) denote the class of a patient (i.e. a patient of class (k, g) is assigned a surgeon g and belongs to DRG group k). Patients belonging to different classes are allowed to differ in terms of interarrival times, service times and routing. Assume interarrival times and service times of patients to be i.i.d. if they belong to one and the same class and assume them to be independently (but not necessarily identically) distributed otherwise. Let $\eta_{i(k,g)}$ denote the external arrival rate of a class (k, g) patient at workstation i (remark that external arrivals are only assumed to take place at the consultation workstations). The aggregate external arrival

rate at a workstation i equals

$$\eta_i = \sum_{k=1}^{K} \sum_{g=1}^{G} \eta_{i(k,g)}. \tag{1}$$

Note that expression 1 is a general expression, most of the time a workstation will be uniquely assigned to a single surgeon, making the summation over g redundant.

We assume that the interarrival times of the external arrivals are exponentially distributed. Such an assumption poses only a slight restriction on the accuracy of the model while it has been shown by Palm [15] and Khinchin [16] that the sum of a large numbers of independent renewal processes (i.e. the arrival processes of the different classes of patients) will tend to a Poisson process. Considering the multitude of classes of patients, the approximation of the aggregate external arrival process by means of a Poisson process should be accurate. In addition, Lariviere et al. [17] show that the assumption of exponential interarrival times is reasonable in many service systems.

Let $\gamma_{i(k,g)}$ denote the expected number of visits a class (k, g) patient will make to workstation i (remark that only the consultation workstations are assumed to be visited more than once). The aggregate arrival rate of patients at the consultation level equals

$$\lambda_i = \sum_{k=1}^{K} \sum_{g=1}^{G} \eta_{i(k,g)} \gamma_{i(k,g)}, \ \forall i \in \{1, 2, \ldots, 6\}. \tag{2}$$

Remark that in contrast to the aggregate external arrival rate, which was assumed to be Poisson-distributed, the aggregate arrival rate (at each of the workstations) is allowed to follow a general distribution. Further define the routing matrix R in which the elements r_{ij} indicate the probability of a patient to travel from station i to station j after service completion at station i. Adhering to standard conventions, we establish a node (of index $i = 0$) from which external arrivals originate and which also serves as a sink for patients leaving the hospital system. Let r_{i0} indicate the probability of leaving the system when departing from station i. Conversely r_{0i} implies the probability of an

external arrival occurring at station i. The probabilities r_{ij} can be expressed as the the proportion of the arrivals at station i that travel towards station j. When assuming the stability of the queueing network, the law of conservation of flows (what comes in, must go out) dictates

$$r_{i0} = \frac{\eta_i}{\lambda_i}. \tag{3}$$

With respect to the surgery workstations, each patient visiting the hospital department is subjected to surgery exactly once. As such, one can infer that

$$\lambda_i = \eta_i,\ \forall i \in \{7, 8, \ldots, 12\}. \tag{4}$$

Hence the probability of transition from the consultation level to the surgery level may be defined as

$$r_{ij} = \frac{\eta_i}{\lambda_i},\ \forall i \in \{1, 2, \ldots, 6\},\ j = i + I. \tag{5}$$

Finally, at the consultation level, the probability of re-entry equals

$$r_{ii} = 1 - (r_{i0} + r_{ij}) = 1 - \frac{2\eta_i}{\lambda_i},\ \forall i \in \{1, 2, \ldots, 6\},\ j = i + I. \tag{6}$$

The routing probabilities of transferring from a surgery workstation $i,\ i \in \{6, 7, \ldots, 12\}$ towards a recovery ward $j,\ j \in \{13, 14, 15\}$ is obtained as follows

$$r_{ij} = \frac{\lambda_j^{(i)}}{\lambda_i},\ \forall i \in \{7, 8, \ldots, 12\},\ \forall j \in \{13, 14, 15\}, \tag{7}$$

where $\lambda_j^{(i)}$ is the empirically observed arrival rate of patients at recovery workstation $j,\ j \in \{13, 14, 15\}$ originating from surgery workstation $i,\ i \in \{6, 7, \ldots, 12\}$. Remark that $\lambda_j^{(i)} = 0,\ \forall i \ni \{6, 7, \ldots, 12\},\ \forall j \ni \{13, 14, 15\}$. As such, the arrival rates at

recovery equal

$$\lambda_j = \sum_{i=0}^{I} \lambda_j^{(i)},\ \forall j \in \{13, 14, 15\}\,. \tag{8}$$

From this we obtain

$$r_{ij} = \frac{\lambda_i^{(j+I)}}{\lambda_i},\ \forall i \in \{13, 14, 15\}\,,\ \forall j \in \{1, 2, \ldots, 6\}\,. \tag{9}$$

All other routing probabilities stem directly from the structure of the model (e.g. the probability of returning to the consultation workstation after the completion of recovery equals unity). The subsequent set of expressions completely defines the routing probabilities:

$$\begin{aligned}
&r_{i0} = \tfrac{\eta_i}{\lambda_i},\\
&r_{ii} = \delta_{ii} - \tfrac{2\eta_i}{\lambda_i},\\
&r_{ij} = \delta_{ij}\tfrac{\eta_i}{\lambda_i},\ \forall i \in \{1, 2, \ldots, 6\}\,,\ i \neq j,\ j > 0,\\
&r_{ij} = \tfrac{\lambda_j^{(i)}}{\lambda_i},\ \forall i \in \{7, 8, \ldots, 12\}\,,\ i \neq j,\ j > 0,\\
&r_{ij} = \tfrac{\lambda_i^{(j+I)}}{\lambda_i};\ \forall i \in \{13, 14, 15\}\,,\ i \neq j,\ j > 0,
\end{aligned} \tag{10}$$

where $(\delta_{ij} = 1)$ if at least one of the patient classes travels from station i to station j and $(\delta_{ij} = 0)$ otherwise.

Remark that other routing structures give rise to other routing probabilities. The routing structure and corresponding equations discussed in this section are only valid under the previously imposed assumptions concerning patient flow.

With respect to the service times, let $f_{i(k,g)}(x)$ denote the natural service time probability density function of a class (k, g) patient visiting workstation i. Have $1/\nu_{i(k,g)}$ and $\sigma^2_{\nu_{i(k,g)}}$ represent the average natural service time for a class (k, g) patient at workstation i and its variance respectively. The natural process time excludes random interruptions, absences and any other external influence. Assume service times of different classes to be independent but not necessarily identically distributed. The probability that a ran-

domly picked unit in front of the workstation is of class (k, g) is given by $\lambda_{i(k,g)}/\lambda_i$, where $\lambda_{i(k,g)}$ is the total arrival rate of class (k, g) patients at workstation i. Define the probability function of the aggregate natural service times at station i as follows

$$f_i(x) = \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{\lambda_{i(k,g)}}{\lambda_i} f_{i(k,g)}(x). \tag{11}$$

As a result the average natural service time requirement of a unit in front of the workstation amounts to

$$\frac{1}{\nu_i} = \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{\lambda_{i(k,g)}}{\lambda_i} \frac{1}{\nu_{i(k,g)}}. \tag{12}$$

When observing the variance of the aggregate natural service process, one can deduce that

$$\begin{aligned} \sigma^2_{\nu_i} &= \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{\lambda_{i(k,g)}}{\lambda_i} \int \left(x - \frac{1}{\nu_i}\right)^2 f_{i(k,g)}(x)\, dx, \\ &= -\frac{1}{\nu_i^2} + \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{\lambda_{i(k,g)}}{\lambda_i} \left(\sigma^2_{\nu_{i(k,g)}} + \frac{1}{\nu^2_{i(k,g)}}\right). \end{aligned} \tag{13}$$

We refer to $\sigma^2_{\nu_i}$ as a measure of the natural variability of the aggregate process times at workstation i. The same result was obtained by Whitt [18] and has widely been adopted in literature (for instance refer to Whitt [19] and Haskose et al. [20]).

3.2 Variability from preemptive and nonpreemptive outages

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 Babe et al. [21], Liu et al. [22], Chisholm et al. [23] and Chisholm et al. [24] 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. [25] 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 [22] and may lead to interruptions during the treatment process (see Chisholm et al. [23], Chisholm et al. [24] and Easton et al. [25]). It is clear that hospital environments are characterized by substantial amounts of variability. As is argued in the literature [26], 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 (see Vissers et al. [5], Vandaele et al. [27] and Sethuraman et al. [28]). 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 et al. [29] and Haque et al. [30]. 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 2. Next to the modeling of planned absences (e.g.

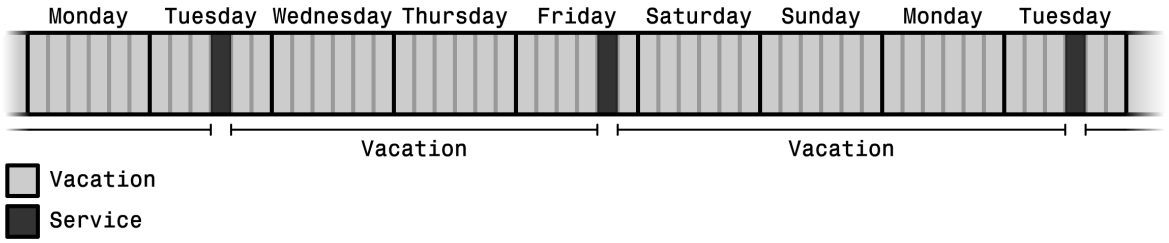

Figure 2: Illustration of a vacation model

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 [31] and Takagi [32]. A more recent yet less general survey can be found in Vishnevskii et al. [33]. 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 Vishnevskii et al. [26]. In their work, Hopp et al. [26] propose a transformation of the service process times to account for service outages. The results of Hopp et al. [26] are widely accepted in the literature (for instance refer to Lambrecht et al. [13]). In this work, we develop new expressions to model the impact of service outages that are peculiar to healthcare systems. In what follows, we first discuss the difference between preemptive and nonpreemptive outages. Next, we provide the means to model them.

3.2.1 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. We call these increased, adjusted service times, effective processing 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 discuss the nonpreemptive outages. Nonpreemptive outages typically occur between jobs, rather than during jobs. They occur at the beginning of each service session (i.e. 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 unplanned absences and define the mean and variance of the amount of time absent as $1/\mu_s$ and σ_s^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 session (e.g. a consultation work shift). Each start of a service session may induce a delay due to an absence. In other words, the number of patients in a service session is a decision variable and is comparable to a lot sizing decision. Evaluating different service session sizes (i.e. different values of n) may provide key managerial insights. We will address this issue in an upcoming section.

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. [26]. They are characterized by a Mean Time To Interrupt (τ_f) and a Mean Time To Resolve (τ_r). The model presented in Hopp et al. [26] presumes interrupts to occur only during actual service time. However, in a hospital setting it is not inconceivable 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 occurrence of nonpreemptive and preemptive outages. In order to maintain transparency of the model and of notation, we impose the following assumptions: (1) service outages only occur at the consultation level (i.e. only workstations i, $i \in \{1, 2, \ldots, 6\}$ are affected); (2) for each of the surgeons, the impact of outages is identical (i.e. $1/\mu_s$, σ_s^2, n, τ_f and τ_r remain the same for each of the workstations at the consultation level).

3.2.2 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 does not depend on the service history (i.e. they are independent of prior services and absence times). The absence times themselves are distributed following a probability density function $f_s(x)$. The average absence time and its variance are represented by $1/\mu_s$ and σ_s^2. 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 3. The probability density function of the combined service times at a workstation i equals

$$f_{i_c}(x+y) = \sum_{k=1}^{K} \sum_{g=1}^{G} \frac{\lambda_{i(k,g)}}{\lambda_i} f_{i(k,g)}(x) f_s(y). \tag{14}$$

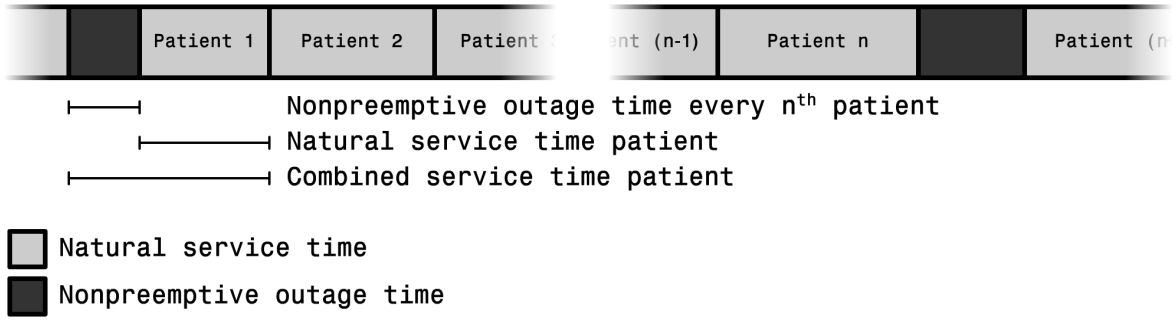

Figure 3: The combined service time

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. The other services as a whole have a probability $((n-1)/n)$ of randomly being picked. Therefore, we can define the mean aggregate service times including the effect of absence times as follows

$$\begin{aligned} \frac{1}{\upsilon_i} &= \left[\left(\frac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i(k,g)}(x)\,x dx\right] + \\ &\quad \left[\frac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i(k,g)}(x)\,f_s(y)\,(x+y)\,dydx\right], \\ &= \frac{1}{\nu_i}+\frac{1}{n\mu_s}. \end{aligned} \tag{15}$$

With respect to the variance of the aggregate service time (including absence times) at the consultation workstations we develop the following expression

$$\begin{aligned} \sigma^2_{\upsilon_i} &= \left[\left(\frac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i(k,g)}(x)\left(x-\frac{1}{\upsilon_i}\right)^2 dx\right] + \\ &\quad \left[\frac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i(k,g)}(x)\,f_s(y)\left(x+y-\frac{1}{\upsilon_i}\right)^2 dydx\right], \\ &= \sigma^2_{\nu_i}+\frac{\sigma_s^2}{n}+\frac{1}{\mu_s^2}\left(\frac{n-1}{n^2}\right). \end{aligned} \tag{16}$$

The above expression is equivalent to that of Hopp et al. [26] and is valid under the assumption that the combined service times as well as ordinary service times are independently distributed.

3.2.3 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_f 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_f), 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 τ_{r0_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 σ_r^2. In addition, resolve times of different outages are assumed to be 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) at a workstation i can be expressed as

$$\frac{1}{\omega_i} = \frac{1}{\nu_i} + \sum_{j=1}^{J_0} \tau_{r0_j}. \tag{17}$$

As such, the average service time $1/\omega_i$ incorporates both the natural service time $1/\nu_i$ 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) and its mean and variance both equal $\left(1/(\nu_i \tau_f)\right)$. We face a sum of random variables (the resolve times τ_{r0_j}) in which the number of random variables (the number of interrupts J_0), is a random variable itself. Assume that J_0 and τ_{r0_j} ($\forall j \in \mathbb{N}$) are i.i.d. variables. In addition assume the mean as well as the variance of τ_{r0_j} to be equal for all $j \in \mathbb{N}$. Therefore, the mean and variance of the sum of J_{i0} random variables τ_{r0_j} can be expressed as [34]

$$E\left[S_0\right] = E\left[J_0\right] E\left[\tau_{r0_j}\right], \tag{18}$$

$$\sigma_{S_0}^2 = E\left[J_0\right] \sigma_r^2 + E\left[\tau_{r0_j}\right]^2 \sigma_{J_0}^2, \tag{19}$$

where S_0 is the random variable representing the sum of J_0 resolve times τ_{r0_j}. In other words we have that

$$S_0 = \sum_{j=1}^{J_0} \tau_{r0_j}. \tag{20}$$

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

$$E\left[S_0\right] = \frac{1}{\nu_i} \frac{\tau_r}{\tau_f}, \tag{21}$$

$$\sigma_{S_0}^2 = \frac{1}{\nu_i} \frac{\sigma_r^2 + \tau_r^2}{\tau_f}. \tag{22}$$

The mean aggregate service time including the effect of interrupts may be expressed as

$$E\left[\frac{1}{\omega_i}\right] = \frac{1}{\nu_i} \frac{\tau_f + \tau_r}{\tau_f}. \tag{23}$$

This corresponds to the expression presented in Hopp et al. [26] 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 approximation of the second moment:

$$E\left[\left(\frac{1}{\omega_i}\right)^2\right] = \left(\sigma_{\nu_i}^2 + \frac{1}{\nu_i^2}\right)\left(1 + \frac{\tau_r}{\tau_f}\right)^2 + \sigma_{S_0}^2. \tag{24}$$

Using the expression for the second moment we obtain the variance of the service times including the effect of interrupts

$$\sigma_{\omega_i}^2 = \sigma_{\nu_i}^2\left(1 + \frac{\tau_r}{\tau_f}\right)^2 + \sigma_{S_0}^2. \tag{25}$$

This expression once more matches the formula derived in Hopp et al. [26]. 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 l ($l \in \mathbb{N}$) denote the set index. We define $\tau_{r_{l_j}}$ to be the resolve time of the j^{th} interrupt belonging to the set of index l (i.e. the interrupt is said to be of order l). 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 l took place during the resolving of interrupts of order $(l-1)$. Figure 4 provides further insight. In addition define S_l as the sum of resolve times corresponding to interrupts of order l. We have that

$$S_l = \sum_{j=0}^{J_l} \tau_{r_{l_j}}, \tag{26}$$

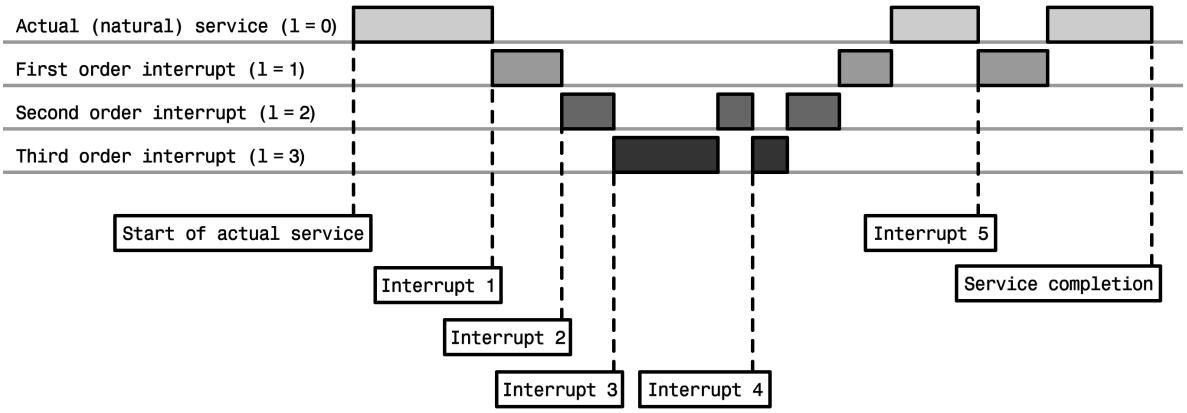

Figure 4: Interrupted service process of a single patient

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

$$E\left[J_l\right] = \sigma_{J_l}^2 = \frac{1}{\nu_i \tau_f}\left(\frac{\tau_r}{\tau_f}\right)^l. \tag{27}$$

One can infer that

$$E\left[S_l\right] = \frac{\tau_r}{\nu_i \tau_f}\left(\frac{\tau_r}{\tau_f}\right)^l, \tag{28}$$

$$\sigma_{S_l}^2 = \frac{\tau_r}{\nu_i \tau_f}\left(\frac{\tau_r}{\tau_f}\right)^l \left(\sigma_r^2 + \tau_r^2\right). \tag{29}$$

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

$$E\left[\frac{1}{\omega_i}\right] = \frac{1}{\nu_i}\frac{\tau_f}{\tau_f - \tau_r}. \tag{30}$$

Using these parameters, the second moment is expressed as

$$E\left[\left(\frac{1}{\omega_i}\right)^2\right] = \left(\sigma_{\nu_i}^2 + \frac{1}{\nu_i^2}\right)\left[1 + 2\frac{\tau_r}{\tau_f - \tau_r} + \left(\frac{\tau_r}{\tau_f - \tau_r}\right)^2\right] + \frac{1}{\nu_i}\frac{\sigma_r^2 + \tau_r^2}{\tau_f - \tau_r}. \tag{31}$$

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

$$\sigma_{\omega_i}^2 = \frac{\tau_f^2 \sigma_{\nu_i}^2 + \frac{1}{\nu_i}\left(\tau_f - \tau_r\right)\left(\sigma_r^2 + \tau_r^2\right)}{\left(\tau_f - \tau_r\right)^2}. \tag{32}$$

3.2.4 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 at a workstation i can be expressed as

$$\begin{aligned} \tfrac{1}{\psi_i} &= \left[\left(\tfrac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i_{f(k,g)}}(x)\,x dx\right] + \\ &\quad \left[\tfrac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i_{f(k,g)}}(x)\,f_s(y)\,(x+y)\,dydx\right], \\ &= \tfrac{1}{\omega_i} + \tfrac{1}{n\mu_s}, \end{aligned} \tag{33}$$

where $f_{i_{f(k,g)}}(x)$ is the probability density function of consultation service times of a class (k,g) patient at a workstation i including the effect of all order interrupts. Its mean and variance are given by $1/\omega_i$ and $\sigma_{\omega_i}^2$ respectively. We refer to $1/\psi_i$ 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 a workstation i may be expressed as

$$\begin{aligned} \sigma_{\psi_i}^2 &= \left[\left(\tfrac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i_{f(k,g)}}(x)\left(x-\tfrac{1}{\psi_i}\right)^2 dx\right] + \\ &\quad \left[\tfrac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i_{f(k,g)}}(x)\,f_s(y)\left(x+y-\tfrac{1}{\psi_i}\right)^2 dydx\right], \\ &= \sigma_{\psi_i}^2 = \sigma_{\omega_i}^2 + \tfrac{\sigma_s^2}{n} + \tfrac{1}{\mu_s^2}\left(\tfrac{n-1}{n^2}\right). \end{aligned} \tag{34}$$

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

3.2.5 Including the time availability of workstations

It is well known that many services do not operate continuously over time. Consultation and surgery typically operate during certain time intervals (service sessions) which means that only a proportion of the total available time can be used effectively. Vacation models are often applied to solve this problem. Another way to handle the problem is to rescale all service processing times so that they fit a preset uniform time scale. In this study we agreed on a 24 hours per day, 7 days per week time scale (basically because this is the appropriate time scale for recovery processes). Let A_i denote the availability of workstation i; A_i represents the available time in proportion to the preset uniform time scale. For instance, if a workstation operates only 6 hours per day, then the availability equals 25%.

When rescaling the service times established in the previous sections, we obtain the total effective service times:

$$\begin{array}{rcll} \frac{1}{\mu_i} & = & \frac{1}{A_i\psi_i}, & \forall i \in \{1,2,\ldots,6\}, \\ \frac{1}{\mu_i} & = & \frac{1}{A_i\nu_i} & \forall i \in \{7,8,\ldots,15\}, \\ \sigma_i^2 & = & \frac{\sigma_{\psi_i}^2}{A_i^2}, & \forall i \in \{1,2,\ldots,6\}, \\ \sigma_i^2 & = & \frac{\sigma_{\nu_i}^2}{A_i^2} & \forall i \in \{7,8,\ldots,15\}. \end{array} \tag{35}$$

The above procedure results in the total effective service times including natural process time, the effect of outages and the impact of availability of workstations. The mean total effective service time and its variance can now be used to compute the squared coefficient of variation

$$C_{s_i}^2 = \sigma_i^2 \mu_i^2. \tag{36}$$

3.2.6 Squared coefficient of variation of the aggregate arrival process

In order to approximate the parameters of the aggregate arrival process, some more challenging arithmetics are needed. It was pointed out by Albin [35] that if at least one of the interarrival time distributions, constituting the arrival process, does not stem from a Poisson process, the resulting aggregate interarrival times do no longer hold the property of independence. As a result the analytical analysis of the aggregate arrival process becomes highly intractable. Therefore approximations will be adopted to assess the variance and, more important, the squared coefficient of variation of the aggregate arrival process. The squared coefficients of variation of the aggregate arrivals at the different workstations will be extracted using a technique which was pioneered by Shanthikumar et al. [10]. This technique implies the use of a set of linear equations which has to be solved in order to obtain the squared coefficients of variation of the arrivals. This approach is widely adopted in literature [36] and was later generalized by Lambrecht et al. [13]. Using the technique that was outlined in Lambrecht et al. [13], we are given a set of I equations:

$$-\sum_{i=1}^{I} \lambda_i r_{ij}^2 (1-\rho_i^2) C_{a_i}^2 + \lambda_j C_{a_j}^2 = \sum_{i=1}^{I} \lambda_i r_{ij} (r_{ij}\rho_i^2 C_{s_i}^2 + 1 - r_{ij}) + \eta_j C_{a_{\eta_j}}^2, \qquad (37)$$

where η_j and $C_{a_{\eta_j}}^2$ denote the rate and squared coefficient of variation of the aggregate external arrival process at station j respectively. In addition, ρ_i represents the effective traffic intensity at workstation i and equals λ_i/μ_i. While all elements except the I squared coefficients of variation are known, we are presented with a system of I equations yielding I unknowns. Solving this set of linear equations provides us with the I unknown squared coefficients of variation (i.e. $C_{a_i}^2; \forall i \in \{1, \ldots, I\}$).

With all model parameters firmly defined, we now have a solid base to carry out the performance evaluation of the hospital department. In the upcoming section we

discuss a numerical example of the model presented above and provide some practical applications.

4 Applications

In this section, we discuss a numerical example using the queueing model described in the previous section. Next, we illustrate the devastating impact of service interruptions on patient flow times. Subsequently we show the potential gains obtained by pooling hospital resources. Finally, we present an optimization model to determine the optimal number of patients to be treated during a service session.

4.1 Numerical example

The numerical example presented in this section builds on data gathered at the orthopedic department of the Middelheim hospital in Belgium. Using these empirical data as inputs, the flow time of patients at the hospital department may be assessed using so-called flow time expressions. A variety of flow time expressions are available in queueing literature. A previous study has shown the Kingman equation [26] to yield accurate results when assessing the flow times of patients in complex hospital systems [4]. As such, in the remainder of this article, we will use the Kingman equation to determine patient flow times. With respect to the Kingman equation, one can define the expected flow time of a patient at workstation i as follows

$$E\left[W_i\right] = \left(\frac{C_{a_i}^2 + C_{s_i}^2}{2}\right)\left(\frac{\rho_i^{\sqrt{2(m_i+1)}-1}}{m_i\left(1-\rho_i\right)}\right)\frac{1}{\mu_i} + \frac{1}{\mu_i}, \tag{38}$$

where m_i denotes the number of parallel servers at workstation i ($m_i = 25\ \forall i \in \{13, 14, 15\}$). If only a single server is present (i.e. at workstations $i,\ i \in \{1, 2, \ldots, 12\}$),

i	1	2	3	4	5	6
$1/\psi_i$	24.85	24.85	24.85	24.85	24.85	24.85
$1/\mu_i$	310.7	690.4	310.7	167.9	155.3	248.5
$C_{s_i}^2$	1.334	1.334	1.334	1.334	1.334	1.334
$1/\lambda_i$	329.8	741.5	317.0	174.5	167.5	268.8
$C_{a_i}^2$	1.026	1.418	1.051	0.759	0.752	0.952
A_i	0.080	0.036	0.080	0.148	0.160	0.100
ρ_i	0.942	0.931	0.980	0.962	0.927	0.925
$E[W_i]$ (days)	4.360	9.402	12.90	3.219	1.547	2.593

Table I: Summary Table of the model results (workstations 1 to 6)

i	7	8	9	10	11	12
$1/\nu_i$	110.0	96.20	89.17	57.50	56.35	93.18
$1/\mu_i$	1048	2004	1351	845.7	593.2	1035
$C_{s_i}^2$	0.266	0.406	0.203	0.171	0.165	0.274
$1/\lambda_i$	1,111	2,111	1,380	883.4	620.5	1,073
$C_{a_i}^2$	1.089	1.121	1.074	1.058	1.068	1.070
A_i	0.105	0.048	0.066	0.068	0.095	0.090
ρ_i	0.943	0.950	0.979	0.957	0.956	0.965
$E[W_i]$ (days)	8.907	21.38	29.42	8.674	5.918	14.14

Table II: Summary Table of the model results (workstations 7 to 12)

no pooling is assumed to take place and the formula reduces to

$$E[W_i] = \left(\frac{C_{a_i}^2 + C_{s_i}^2}{2}\right)\left(\frac{\rho_i}{1-\rho_i}\right)\frac{1}{\mu_i} + \frac{1}{\mu_i}. \tag{39}$$

Using the empirical data, resulting flow times at each of the workstations are obtained. The results are presented in Table I and Table II (all results are expressed in minutes unless indicated otherwise). While no waiting occurs at the wards (i.e. the process of recovery takes place immediately after surgery) the performance measures of workstations 13 to 15 are not included here. With respect to consultation, no distinction was made between the different surgeons. One can observe that the effective service time

(including the effect of interrupts and absences) amounts to 24.85 minutes (the natural service time amounting to 15 minutes). The coefficient of variation equals 1.334 (the natural coefficient of variation amounting to 0.6386). Arrival rates and their variances depend on the number of patients visiting each surgeon. The utilization rates of the surgeons are all very high, which translates into significant average patient flow times varying from 1.5 days to 12.9 days.

Similar observations may be made with respect to surgery. Here we allow surgeons to have different processing times depending on the type of surgery they perform. In addition, observe the significantly longer flow times for patients at the surgery level.

4.2 The impact of interrupts

The impact of interrupts on medical practice has been observed by Harvey et al. [37], Lehaney et al. [38], Chisholm et al. [24], Brixey et al. [39], France et al. [40], Volpp et al. [41], Tucker et al. [42] and Gabow et al. [43] among others. All agree on the detrimental effects of interrupts on patient flow time. In order to demonstrate these detrimental effects, we present a number of scenarios in which we gradually reduce the impact of interrupts. We build on the setting of the hospital department discussed previously. To maintain transparency, we focus on a single consultation workstation (i.e. the only workstations that are susceptible to interrupts during the service process). We adjust the mean time to interrupt (i.e. τ_f) at this workstation to assess the varying impact of interrupts (all other model parameters remain unchanged). The results are given in Table III. Note that we used the third workstation to study the impact of various degrees of interrupts (the results corresponding to the numerical example presented in section 4.1 are indicated in bold). Figure 5 illustrates the phenomenon graphically. It is clear that heavy traffic systems (i.e. systems which operate under high workload) benefit greatly from even a small reduction in utilization rate. Unfortunately, only

τ_f	$E[W]$	ρ	τ_f	$E[W]$	ρ	τ_f	$E[W]$	ρ
10.4	183.2	0.998	11.6	16.24	0.984	18	4.433	0.943
10.5	93.58	0.997	11.8	14.35	0.982	20	3.393	0.936
10.6	63.28	0.995	**12.0**	**12.90**	**0.980**	25	3.288	0.924
10.7	48.05	0.994	12.5	10.43	0.975	30	2.968	0.916
10.8	38.88	0.993	13.0	8.880	0.971	40	2.652	0.907
10.9	32.76	0.992	14.0	7.029	0.963	60	2.401	0.897
11.0	28.38	0.990	15.0	5.966	0.957	80	2.294	0.893
11.2	22.54	0.988	16.0	5.276	0.952			
11.4	18.82	0.986	17.0	4.791	0.947			

Table III: Impact of interrupts (expressed in minutes) on patient flow time (expressed in days) at a single workstation

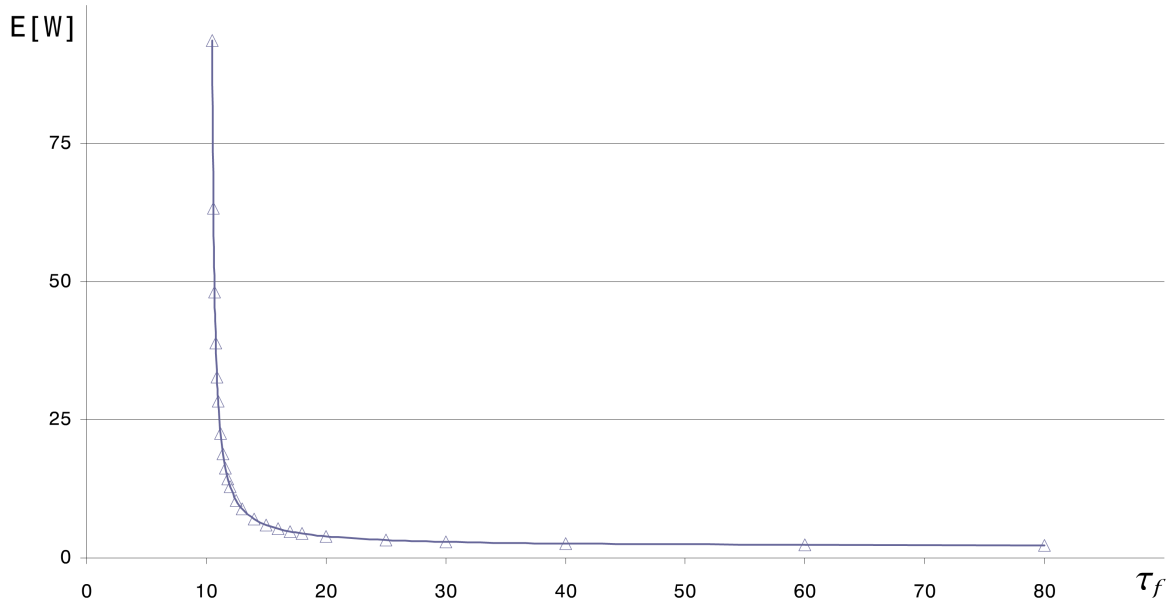

Figure 5: Varying impact of interrupts (expressed in minutes) and the effect on patient flow times (expressed in days)

limited means are available to achieve such a reduction in utilization rate. A variety of options arise:

- The most obvious way to reduce the effective utilization is process improvement. Continuous improvement and six sigma programs are very beneficial. Reducing the frequency of interrupts can be classified in this category.

- Expand capacity; hospital resources such as operating theatres, scanners and other equipment are often operating at maximum capacity. Expanding capacity would be an effective means to reduce hospital workload. However, expanding capacity is often very expensive or is simply impossible (e.g. due to legal constraints).

- Limit patient volumes; a reduction in hospital workload might also be achieved by limiting the amount of patients receiving treatment. Pursuing this option however, results in loss of hospital income and a reduced level of service.

In literature, valuable insights are provided that offer guidance in the quest to reduce the impact of interrupts. For instance, Harvey et al. [37] suggest the pooling of paging of doctors (next to telephone calls, paging calls are one of the largest sources of interrupts) in order to decrease variability in individual paging patterns. France et al. [40] propose the use of information systems (e.g. an electronic whiteboard) and team training to enhance performance. Tucker et al. [42] suggest the redesign of treatment processes (e.g. outsourcing of administrative tasks) in order to make service more robust against preemptive outages. In addition Tucker et al. [42] and Volpp et al. [41] propose the filtering of non-urgent communication towards medical staff. These and other practical guidelines enable hospital decision makers to minimize the impact of interrupts on the service process.

4.3 The impact of pooling

Pooling refers to the aggregation (consolidation) of the demand from multiple items into one, such that the consolidated demand can be satisfied from a single buffer. More specifically, capacity pooling refers to the idea of sharing available capacity among various sources of demand (e.g. patient classes). In a hospital setting this refers to the sharing of expensive diagnostic equipment, wards or labs. In a non-pooling environment, each resource fulfills its own demand, relying solely on its own capacity. In a pooled environment, demand is aggregated and fulfilled from a single shared facility. A rich literature on pooling in queueing systems exists. For an excellent overview, refer to Benjaafar et al. [44] and Yu et al. [45].

It has long been know that pooling is beneficial to system performance. More specifically, pooling allows to maintain a specified level of service quality (e.g. patient flow times) with less capacity requirements. The beneficial effect of pooling stems from the increased ability of the system to cope with variability. For instance, in pooled systems, it is much less likely for the queue to be empty. As such, the impact of variability in the arrival pattern of patients (patients may arrive early, late or may even fail to show up at all) or in the service process of surgeons is minimized.

In this section, we demonstrate the impact of capacity pooling by means of a small experiment. We build on the setting of the hospital department discussed in the previous sections. In the experiment the servers at the consultation and surgery level are pooled. The following assumptions are imposed:

- Patients are treated by the first surgeon available for service, even if the patient was assigned another surgeon upon first arrival at the hospital.

- Surgeon working schedules are identical and no structural constraints are imposed (i.e. it should be possible to service 6 patients simultaneously).

i	1	2
$1/\mu_i$	246.90	995.87
$C_{s_i}^2$	1.334	0.224
$1/\lambda_i$	43.56	173.2
$C_{a_i}^2$	0.996	1.075
A_i	0.101	0.079
ρ_i	0.944	0.958
$E[W_i]$ (pooled)	0.518	1.612
$E[W_i]$ (non-pooled)	4.523	12.47

Table IV: Summary table of the model results after pooling (consultation and surgery workstations)

Returning to our example setting, the six consultation and the six surgery workstations are replaced by a single consultation and a single surgery workstation respectively. Each of these workstations has six parallel servers in operation. The resulting queueing network contains five workstations i, $i \in \{1, 2, \ldots, 5\}$. Let station 1 to 5 represent consultation, surgery, day hospital, internal ward and external ward respectively. When retaining all other characteristics of the setting discussed in the previous sections, one can use the Kingman equation to obtain patient flow times. The resulting performance measures are presented in Table IV (the non-pooled flow times are the weighted average of the flow times observed at the consultation and surgery workstations presented in section 4.1).

The benefits of pooling are clear. Without increasing capacity or altering any of the other system characteristics (except of course the pooling of capacity) we are able to reduce patient flow times at the consultation and surgery level by a factor of 8.73 and 7.74 respectively. Unfortunately, it is often impossible to achieve such a high degree of pooling in a real life hospital system. One quickly runs into a number of limitations:

- Unique relation between patient and surgeon; patients will often refuse to consult another surgeon.

- Limited flexibility of resources; each surgeon has his own specialization. It is often impossible, even for surgeons at the same department, to pass on jobs. As such, the flexibility of surgeons is limited.

- Resources often operate at different time instances; for pooling to take place surgeons need to operate at the same time instance. Due to busy schedules and other limitations, this is not always possible.

- Structural characteristics may further limit the practical applicability of pooling. For instance, if only two operating theatres are available, it is impossible to pool the capacity of the six surgeons at the surgery level. In other words, the bottleneck has shifted from the surgeons onto the number of available operating theatres.

Notwithstanding these constraints, it should be clear that even small amounts of pooling may yield significant reductions in patient flow time. Therefore the pooling of hospital resources is a worthwhile matter for further investigation.

4.4 Finding the optimal number of patients in a service session

The impact of absences at the start of a consultation or surgery session is discussed in Babe et al. [21], Liu et al. [22], Liu et al. [46] and Easton et al. [25]. There is a general agreement on the disruptive effect of absences on patient flow time. Easton et al. [25] identify robust staffing, scheduling and recovery practices to minimize the effects of absences. Liu et al. [46] acknowledge the importance of consultation and surgery block size (i.e. the number of patients treated during a consultation session) and propose a what-if simulation approach in order to determine the best block size. In fact, the relationship between block size and patient flow time is akin to the relationship between batch size and waiting time (in the presence of setups between batches in a manufacturing setting). As such the convex relationship first described by Karmarkar

[47] may also be observed here. In this view, Vandaele et al. [48] determine the optimal size of patient groups queueing in front of a nuclear resonance scanner. We build on the model of Lambrecht et al. [49] in order to determine the optimal number of patients that receives treatment during a service session.

Two conflicting effects may be observed:

- The grouping effect; referring to the time required to assemble a batch of size n. The larger the batch size, the longer patients will have to wait before receiving service.

- The saturation effect; the smaller the batch size, the more service sessions are initiated, the larger the probability of having an absence of medical staff at the start of a service session.

We illustrate these effects in Figure 6. The combination of both effects results in a convex relationship, which implies that there is an optimal group size minimizing average patient flow time. In what follows, we develop the mathematical model to address the batch size decision problem. The objective is to determine the batch size that minimizes the average patient flow time.

In this section we build on the third workstation discussed in the base case (other workstations at the consultation and surgery level may also be analyzed in a similar fashion). To maintain the transparency of the model, we omit the index i referring to the original workstation used in this experiment. Other than the batching of patients, the dynamics of the workstation remain unchanged (as compared to the numerical example presented in section 4.1).

Once sufficient patients are available, a batch (i.e. the equivalent of a service session workload) is created and is introduced into a queue (it is clear that this grouping does not imply that patients have to wait physically in the hospital). Whenever the server

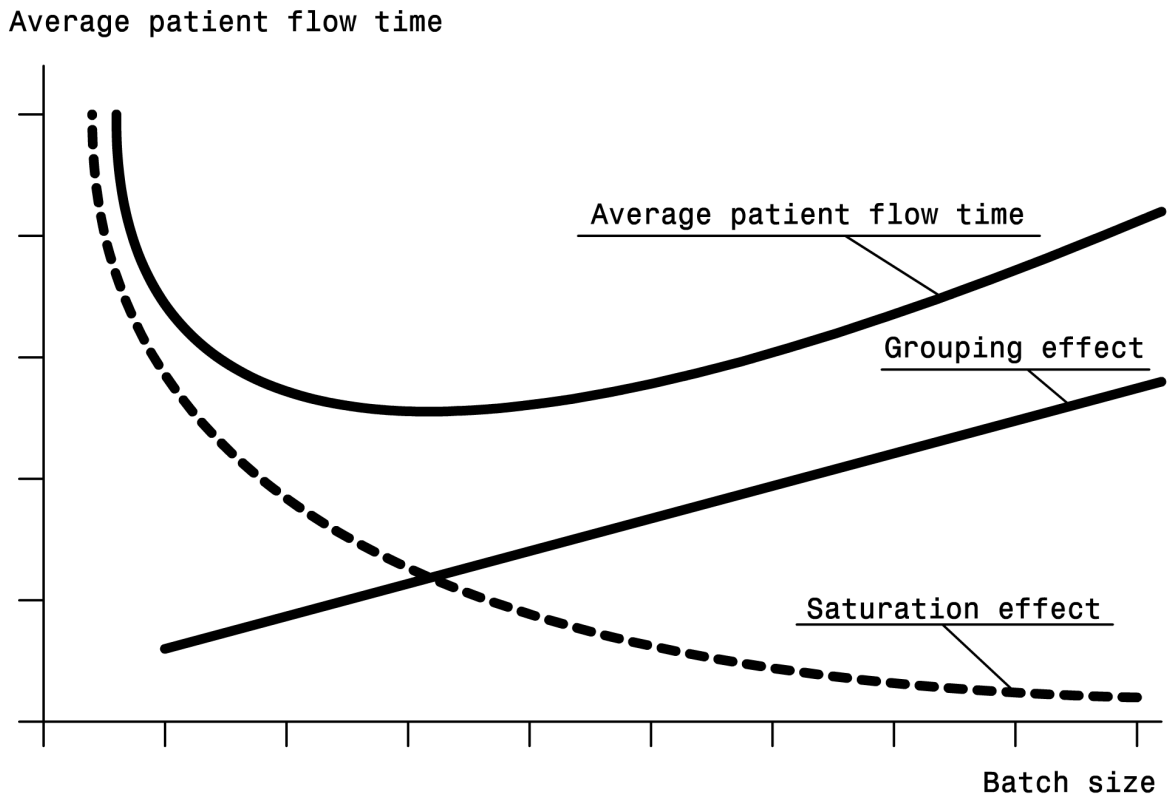

Figure 6: Convex relationship between average patient flow time and batch size

is idle, the batch as a whole receives service. After service, the batch is separated and patients resume their individual routings. A batch of patients is characterized by:

- a batch size n,
- a batch arrival rate λ_b,
- a coefficient of variation of the interarrival times of the batches $C_{a_b}^2$,
- a batch service rate μ_b,
- a coefficient of variation of the service times of the batches $C_{e_b}^2$,

where

$$\begin{aligned} \lambda_b &= n\lambda, \\ C_{a_b}^2 &= \tfrac{C_a^2}{n}, \\ \mu_b &= n\mu, \\ C_{e_b}^2 &= \tfrac{C_e^2}{n} \end{aligned} \tag{40}$$

and λ, C_a^2, μ, C_e^2 are the respective arrival rate, the squared coefficient of variation of the interarrival times, the service rate and the squared coefficient of variation of the service times of the individual patients visiting the third workstation.

The flow time of a patient in this system contains the following elements:

- The collection time; the time required until sufficient patients have arrived and a batch may be processed. The larger the batch size, the longer it takes to gather sufficient patients in order to perform a batch service.
- The waiting time of the batch itself; other batches (i.e. service sessions) may have to be serviced first.
- The absence time; prior to the service of a batch of patients, there exists a probability that the surgeon (or another crucial hospital resource) is absent. The batch

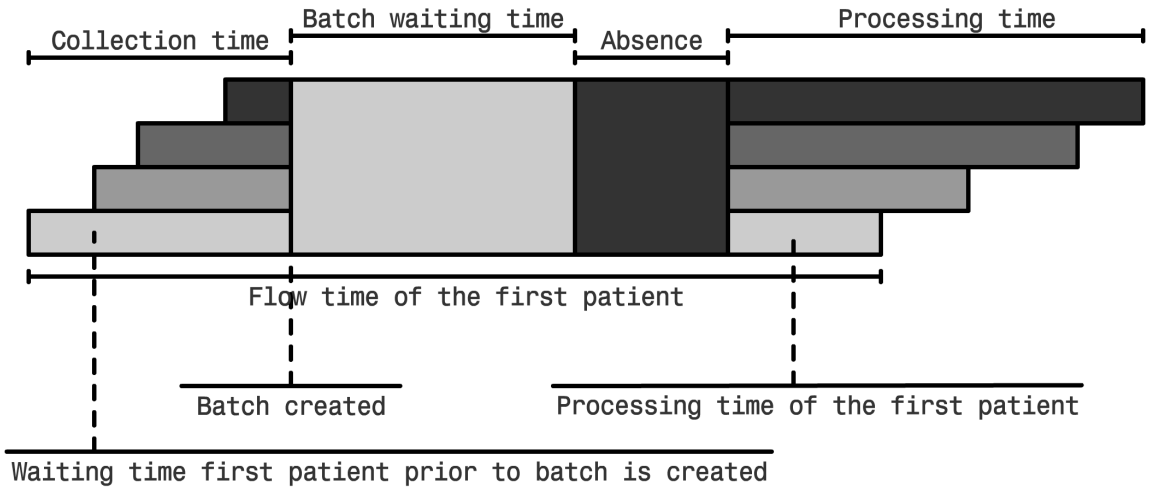

Figure 7: Visualization of the different phases of the batch flow time

of patients has to wait for the surgeon in order to receive service. This absence time can be considered as a setup time for the batch.

- The actual processing of individual patients in the batch.

We visualize the flow time of a patient in Figure 7. The expected flow time of a single patient in the system can be expressed as [49]:

$$E[W] = \frac{n-1}{2\lambda} + E[W_q] + \frac{1}{\mu_s} + \frac{n+1}{2\mu}. \tag{41}$$

This flow time clearly consists of four building blocks. The first term corresponds to the average time a patient will have to wait until a group of size n has been formed (i.e. the collection time). The term $E[W_q]$ stands for the average time that a batch of patients spends waiting in queue until the server becomes idle. We approximate $E[W_q]$ by means of the Kingman equation and obtain

$$E[W_q] = \left(\frac{C_{a_b}^2 + C_{s_b}^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu_b}, \tag{42}$$

where ρ is the effective utilization rate at the third workstation and is given by [49]:

$$\rho = \frac{n\lambda}{n\mu + \mu_s}. \tag{43}$$

The third term corresponds to the absence time that is incurred at the start of a service session in which a batch of patients receives treatment. Both the second and third term are the same for all patients in the batch. The last term indicates how much time a patient spends on processing itself. At this point the model is complete and we can formally state our optimization problem:

$$\begin{aligned} &\text{Minimize } E[W], && E[W] = \tfrac{n-1}{2\lambda} + E[W_q] + \tfrac{1}{\mu_s} + \tfrac{n+1}{2\mu}, \\ &s.t. && \rho < 1, \\ & && n \geq 1. \end{aligned}$$

When using the setting of the hospital department outlined in the previous sections, we are able to provide a numerical example. To maintain transparency, we select a single consultation workstation and assess different values of n in order to obtain the optimal number of patients to be treated during a service session. A summary of the resulting figures is presented in Table V. An illustration is provided in Figure 8. One can deduce that, for this particular workstation, the optimum is reached when treating 8 patients during each service session. More precisely, given a set of input parameters (absence probability, service- and interarrival times, ...) we are able to determine the optimal number of patients to be treated during a service session.

n	$1/\mu_b$	$C_{e_b}^2$	ρ	$E[W]$
3	82.063	0.2276	1.0787	NA
4	99.418	0.1707	0.9802	27.460
5	116.77	0.1365	0.9210	8.2226
6	134.13	0.1138	0.8815	6.3769
7	151.48	0.0975	0.8534	5.8782
8	168.84	0.0853	0.8322	5.7761
9	186.19	0.0758	0.8162	5.8441
10	203.54	0.0683	0.8027	6.0004
11	220.90	0.0621	0.7919	6.2086
12	238.25	0.0569	0.7830	6.4497
13	255.61	0.0525	0.7754	6.7132
14	272.96	0.0488	0.7689	6.9924
15	290.32	0.0455	0.7632	7.2831
16	307.67	0.0427	0.7583	7.5826
17	325.03	0.0402	0.7540	7.8888
18	342.38	0.0379	0.7501	8.2004
19	359.73	0.0359	0.7466	8.5162
20	377.09	0.0341	0.7435	8.8355

Table V: Summary table of the model results featuring different batch sizes

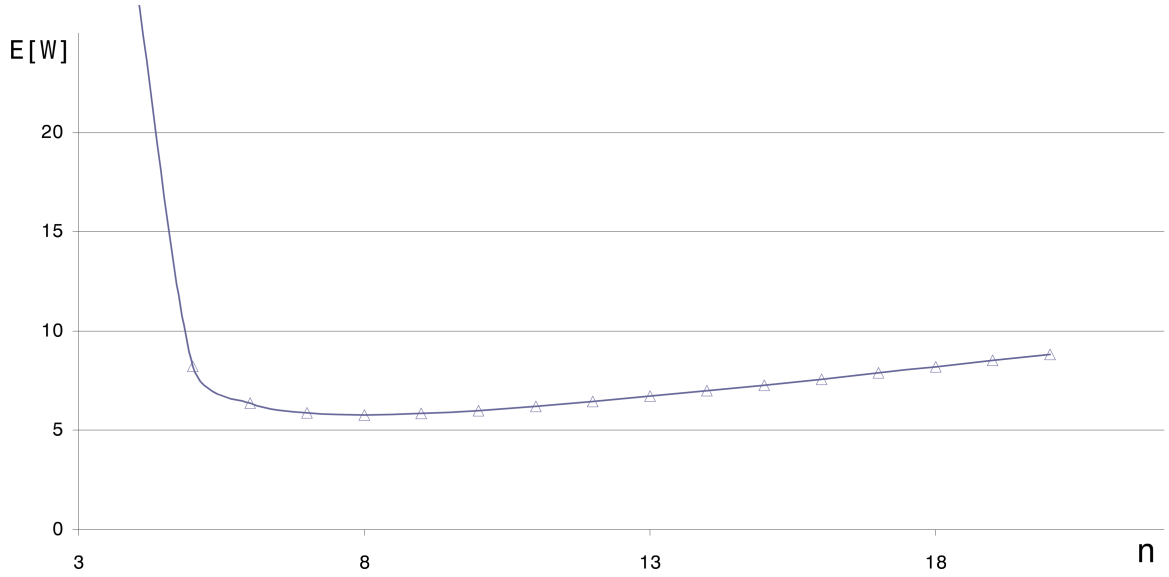

Figure 8: Finding the optimal number of patients

5 Conclusion

In this article we discuss some of the features that differ when modeling healthcare queueing models on the one hand and traditional manufacturing models on the other hand. We show how to implement these features in a hospital queueing system. In addition, we develop new expressions to model service outages that are typical in services in general and in healthcare in particular. The resulting queueing model is used to construct a numerical example and to illustrate a number of practical applications. First we demonstrate the detrimental effect of service interrupts on patient flow times. Next, the beneficial effect of pooling hospital resources is illustrated. Finally, we develop an optimization model that is able to determine the optimal number of patients treated during a single service session.

Notwithstanding these accomplishments, there is still room for improvement. More specifically, improvements may be made with respect to the modeling of time in queueing systems. Open problems include the modeling of time-dependent demand rates, increasing workload as waiting times increase (patients need to be monitored, receive care, ...), ... Moreover, given the inherent high degree of variability in service times, hospitals often use flexible working schedules that allow for overtime, variable server capacity and other deviations from the standard queueing model topology. Such deviations add to the complexity of the problem, making "time" a major modeling issue.

References

[1] R. Hall, D. Belson, P. Muralli, and M. Dessouky, *Modeling patient flows through the healthcare system*, pp. 1–44. Springer Science, New York, 2006. In R.W. Hall: Patient flow: reducing delay in healthcare delivery.

[2] R. Hall, "Patient flow: The new queueing theory for healthcare," *OR/MS Today*, vol. 23, pp. 36–40, 2006.

[3] L. Green and J. Soares, "Computing time-dependent waiting time probabilities in M(t)/M/s(t) queueing systems," *M&SOM Manufacturing & Service Operations Management*, vol. 9, pp. 54–61, 2007.

[4] S. Creemers and M. Lambrecht, "Modeling a healthcare system as a queueing network: the case of a Belgian hospital," Tech. Rep. 0710, Department of Decision Sciences & Information Management, Research Center for Operations Management, Katholieke Universiteit Leuven, 2007.

[5] J. Vissers, J. Bertrand, and G. De Vries, "A framework for production control in health care organizations," *Production Planning & Control*, vol. 12, pp. 591–604, 2001.

[6] A. Roth and R. Van Dierdonck, "Hospital resource planning: concepts, feasibility and framework," *Production and Operations Meanagement*, vol. 4, pp. 2–29, 1995.

[7] G. van Merode, S. Groothuis, and A. Hasman, "Enterprise resource planning for hospitals," *International Journal of Medical Informatics*, vol. 73, pp. 493–501, 2004.

[8] J. Jackson, "Network of waiting lines," *Operations Research*, vol. 5, pp. 518–521, 1957.

[9] J. Kingman, "On queues in heavy traffic," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 24, pp. 383–392, 1962.

[10] J. Shanthikumar and J. Buzacott, "Open queueing network models of dynamic job shops," *International Journal of Production Research*, vol. 19, pp. 255–266, 1981.

[11] G. Bitran and D. Tirupati, "Multiproduct queueing networks with deterministic routing: decomposition approach and the notion of interference," *Management Science*, vol. 34, pp. 75–100, 1988.

[12] W. Whitt, "Approximations for the GI/G/m queue," *Productions and Operations Management*, vol. 2, pp. 114–161, 1993.

[13] M. Lambrecht, P. Ivens, and N. Vandaele, "Aclips: a capacity and lead time integrated procedure for scheduling," *Management Science*, vol. 44, pp. 1548–1561, 1998.

[14] N. Vandaele, L. De Boeck, and D. Callewier, "An open queueing network for lead time analysis," *IIE Transactions*, vol. 34, pp. 1–9, 2002.

[15] C. Palm, "Intensittsschwankungen im fernsprechverkehr," *Ericsson Technics*, vol. 44, pp. 1–89, 1943.

[16] A. Khinchin, *Mathematical Methods in the Theory of Queueing*. Hafner, New York, 1960.

[17] M. Lariviere and J. Van Mieghem, "Strategically seeking service: how competititon can generate poisson arrivals," *Manufacturing & Service Operations Management*, vol. 6, pp. 23–40, 2004.

[18] W. Whitt, "The queueing network analyzer," *The Bell System Technical Journal*, vol. 62, pp. 2779–2815, 1983.

[19] W. Whitt, "Partitioning customers into service groups," *Management Science*, vol. 45, pp. 1579–1592, 1999.

[20] A. Haskose, B. Kingsman, and D. Worthington, "Modelling flow and jobbing shops as a queueing network for workload control," *International Journal of Production Economics*, vol. 78, pp. 271–285, 2002.

[21] M. Babes and G. Sarma, "Out-patient queues at the Ibn-Rochd health center," *Journal of the Operational Research Society*, vol. 42, pp. 845–855, 1991.

[22] L. Liu and X. Liu, "Block appointment systems for outpatient clinics with multiple doctors," *The Journal of the Operational Research Society*, vol. 49, pp. 1254–1259, 1998.

[23] C. Chisholm, E. Collison, D. Nelson, and W. Cordell, "Emergency department workplace interruptions: are emergency physicians "interrupt-driven" and "multitasking"," *Academic Emergency Medicine*, vol. 7, pp. 1239–1243, 2000.

[24] C. Chisholm, A. Dornfeld, D. Nelson, and W. Cordell, "Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices," *Annals of Emergency Medicine*, vol. 38, pp. 146–151, 2001.

[25] F. Easton and J. Goodale, "Schedule recovery: unplanned absences in service operations," *Decision Sciences*, vol. 36, pp. 459–488, 2005.

[26] W. Hopp and L. Spearman, *Factory Physics*. McGraw-Hill Higher Education, New York, 2 ed., 2000.

[27] N. Vandaele and L. De Boeck, "Advanced resource planning," *Robotics and Computer Integrated Manufacturing*, vol. 19, pp. 211–218, 2003.

[28] K. Sethuraman and D. Tirupati, "Evidence of bullwhip effect in healthcare sector: causes, consequences and cures," *International Journal of Services and Operations Management*, vol. 1, no. 4, pp. 372–394, 2005.

[29] K. Stecke and J. Aronson, "Review of operator/machine interference models," *Journal of Production Research*, vol. 23, pp. 129–151, 1985.

[30] L. Haque and M. Armstrong, "A survey of the machine interference problem," *European Journal of Operational Research*, vol. 179, pp. 469–482, 2007.

[31] B. Doshi, "Queueing systems with vacations - a survey," *Queueing Systems*, vol. 1, pp. 29–66, 1986.

[32] H. Takagi, "Queueing analysis of polling models," *ACM Computing Surveys*, vol. 20, pp. 5–28, 1988.

[33] V. Vishnevskii and O. Semenova, "Mathematical methods to study the polling systems," *Automation and Remote Control*, vol. 2, pp. 3–56, 2006.

[34] E. Dudewicz and S. Mishra, *Modern mathematical statistics*. John Wiley Sons, New York, 1988.

[35] S. Albin, "Approximating a point process by a renewal process, II: superposition arrival processes to queues," *Operations Research*, vol. 32, pp. 1133–1162, 1984.

[36] R. Askin, *Modeling and analysis of manufacturing systems*. Wiley, New York, 1993.

[37] R. Harvey, P. Jarrett, and K. Peltekian, "Patterns of paging medical interns during night calls at two teaching hospitals," *Canadian Medical Association Journal*, vol. 151, pp. 307–311, 1994.

[38] B. Lehaney, S. Clarke, and R. Paul, "A case of intervention in an outpatient department," *Journal of the Operational Research Society*, vol. 50, pp. 877–891, 1999.

[39] J. Brixey, M. Walji, J. Zhang, T. Johnson, and J. Turley, "Proposing a taxonomy and model of interruption," in *Proceedings of 6th International Workshop on Enterprise Networking and Computing in Healthcare Industry* (K. Kurokawa, I. Nakajima, and Y. Ishibashi, eds.), pp. 184–188, Healthcom, 2004.

[40] D. France, S. Levin, R. Hemphill, K. Chen, D. Rickard, R. Makowski, I. Jones, and D. Aronsky, "Emergency physicians' behaviors and workload in the precense of an electronic whiteboard," *International Journal of Medical Informatics*, vol. 74, pp. 827–837, 2005.

[41] K. Volpp and D. Grande, "Residents' suggestions for reducing errors in teaching hospitals," *The New England Journal of Medicine*, vol. 348, pp. 851–855, 2006.

[42] A. Tucker and S. Spear, "Operational failures and interruptions in hospital nursing," *Health Services Research*, vol. 41, pp. 643–662, 2006.

[43] P. Gabow, A. Karkhanis, A. Knight, P. Dixon, S. Eiser, and R. Albert, "Observations of residents' work activities for 24 consecutive hours: implications for workflow redesign," *Academic Medecine*, vol. 81, pp. 766–775, 2006.

[44] S. Benjaafar and W. Cooper, "On the benefits of pooling in production-inventory systems," *Management Science*, vol. 51, pp. 548–565, 2005.

[45] Y. Yu and S. Benjaafar, "On service capacity pooling and cost sharing among independent firms," tech. rep., Department of Mechanical Engineering, University of Minnesota, 2006.

[46] L. Liu and X. Liu, "Dynamic and static job allocation for multi-server systems," *IIE Transactions*, vol. 30, pp. 845–854, 1998.

[47] U. Karmarkar, "Lot sizes, lead times and in-process inventories," *Management Science*, vol. 33, pp. 409–418, 1987.

[48] N. Vandaele, I. Van Nieuwenhuyse, and S. Cupers, "Optimal grouping for a nuclear magnetic resonance scanner by means of an open queueing model," *European Journal of Operational Research*, vol. 151, pp. 181–192, 2003.

[49] M. Lambrecht and N. Vandaele, "A general approximation for the single product lot sizing model with queuing delays," *European Journal of Operational Research*, vol. 95, pp. 73–88, 1996.