Online appendix

to

Time-dependent analysis for refused admissions in clinical wards

R. Bekker[a] and A.M. de Bruin[b,a]

[a] Department of Mathematics
VU University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

[b] VU university medical center
De Boelelaan 1117, 1007 MB Amsterdam, The Netherlands

Corresponding author:
René Bekker
Email: rbekker@few.vu.nl
Telephone: +31 (0)20 5982776
Fax: +31 (0)20 5987653

Abstract

For capacity planning issues in health care, such as the allocation of hospital beds, the admissions rate of patients is commonly assumed to be constant over time. In addition to the purely random fluctuations, there is also typically a predictable pattern in the number of arriving patients. For example, roughly 2/3 of the admitted patients at an Intensive Care Unit arrives during office hours. Also, most of the scheduled admissions occur during weekdays instead of during the weekend.

Using approximations based on the infinite-server queue and simulations, we analyze an $M_t/H/s/s$ model to determine the impact of the time-dependent arrival pattern on the required number of operational beds and fraction of refused admissions for clinical wards. In particular, the results show that the effect of the daily pattern is rather limited for clinical wards in contrast to the week-weekend pattern, for which the difference in the fraction of refused admissions across the week is considerable. We also show that an increased variability in length of stay distribution has a stabilizing effect on the time-dependent required number of beds. In the theoretical case of a deterministic length of stay, the variation in refused admissions across the week may be unexpectedly large. Finally, we demonstrate a method to determine the required number of beds across the week.

Keywords: Hospital capacity planning; daily and weekly admission patterns; time-dependent arrivals; refused admissions; infinite-server queues; modified-offered-load approximation.

1 Introduction

Main paper

This is a supplement to the main paper, Bekker and de Bruin (2010), Time-dependent analysis for refused admissions in clinical wards, that appeared in *Annals of Operations Research* **178**, 45–65. Here, we additionally report on the simulation results corresponding to the approximations, see Subsection 5.6 and Appendix B in particular.

The subject of bed capacity planning for health care facilities has already received quite some attention in the operations research literature. Such capacity issues are commonly addressed by using either simulation or queueing models. Within the queueing literature on bed capacity in health care, the $M/M/s$ queue is often applied. This model assumes a Poisson arrival process of patients, exponentially distributed lengths of stay, s operational beds, and an infinitely large waiting room for patients that are waiting until a bed becomes available. Some recent studies [4, 24] show that the $M/G/s/s$ queue (or Erlang loss model) often accurately describes the number of patients present at a ward. In the $M/G/s/s$ model, there is no waiting room for patients. Patients that arrive when all beds are occupied are blocked and thus not admitted to this clinical ward.

A severe restriction for the queueing models mentioned above, is that patients are assumed to arrive uniformly over time. It has been widely recognized that this will typically not be the case for most clinical wards. For instance, across a single day there is an evident peak in the arrival rate during office hours, see e.g. [3, 14]. Furthermore, a structural pattern exists for the number of admissions across the week; the average number of admissions (per day) during the weekdays will generally be larger than the average number of admissions (per day) during the weekends, in particular if the proportion of scheduled arrivals is large. Depending on the operating days of surgical specialists, there may also be a structural difference in the number of admissions for every weekday.

The main goal of this paper is to analyze the impact of a predictable patient arrival pattern on the performance and bed capacity requirements of a clinical ward. To do this, we assume that patients arrive according to a time-dependent Poisson process. We employ a loss model based on [4] in which patients who find all beds occupied are refused. One of our main interests is in the impact of the weekend on the bed occupancy and proportion of refused admissions across the week. Other issues concern the effect of structural fluctuations across the day, the operating room schedule, and the variability of the length of stay distribution. Finally, to exploit the predictable fluctuations in admissions, we consider the case in which the number of operational beds is more flexible.

There already exists a vast body of literature on queueing models with a time-dependent Poisson arrival process. In case the number of servers (beds) is infinite, tractable expressions for the distribution of the number of patients exist, see e.g. [6, 7]. For a finite number of beds, closed-form results are considerably more difficult; results commonly involve numerical results [5, 11, 12] or approximations [12, 23] (see [7] for an overview of approximations based on infinite-server queues).

The effect of time-varying arrivals on the system performance has also been addressed in the literature. In [11, 12], various effects are considered for the case of multi-server Markovian queues with infinite waiting room capacity and the arrival rate following a sine function. These effects include the amplitude of the arrival rate function and the frequency of events. In [5], the sensitivity to the service time distribution of the blocking probability

in a non-stationary loss model is investigated, mainly focusing on a sine function for the arrival rate again. The quality of various approximations for the offered load are considered in [7] for sinusoidal arrival rates. We also refer to [9, 14] for the application of such queueing models for determining staffing requirements in service systems.
Although there are various papers on multi-server queues with non-stationary Poisson arrivals, the impact on the performance for clinical wards has not yet been structurally studied. For example, the assumption of a sinusoidal arrival rate makes it difficult to model the weekend effect. Also, the implications of the variability of the length of stay for clinical wards is not yet fully understood. In this paper, we analyze the performance of an $M_t/H/s/s$ queueing model (where H is the acronym for the hyperexponential distribution), where the arrival rate is piecewise constant. This choice of arrival rate pattern allows for both a flexible set-up as well as tractable approximating results, which can be easily calculated using a spreadsheet. The performance is determined using approximations based on the infinite-server queue (see [23] for a rigorous basis for this type of approximation) and simulation. The approximations give valuable insights into the offered load and the variability and structural pattern of the fraction of refused admissions. They can also be directly applied to determine the time-dependent number of beds, for which the fraction of refused admissions remains relatively stable. For (hyper)exponentially distributed length of stay, which is of prime practical relevance, the approximations perform reasonably well. In the theoretical case of deterministic length of stay, the variation in fraction of refused admission may become unexpectedly large.
This paper is organized as follows. The loss (or blocking) model for clinical wards is described in Section 2 along with the determination of the length of stay distribution. Also, in Subsection 2.1 we describe a basic clinical ward that serves as a reference example for numerical results throughout the paper. In Section 3, some preliminary results regarding infinite server queues, (modified-offered-load) approximations, and square-root staffing formulas are presented. The analysis of the $M_t/H/s/s$ approximation, where the arrival rate is piecewise constant, can be found in Section 4. The impact of varying arrival rates, based on approximations, is presented in Section 5 and includes: predictable fluctuations across the day and week, the effect of different operating days, the impact of the length of stay distribution, and a method to determine the required number of beds across the week. Moreover, the $M_t/D/s/s$ queue is addressed, which is mainly of theoretical relevance. The paper is concluded in Section 6.

2 Model and data analysis

In this paper, we consider the following structural model for the patient flow through a single clinical ward. Patients randomly arrive at the ward. These patients can either be scheduled (elective) or unscheduled (urgent/emergent). A patient is only admitted if an operational bed is available and refused otherwise. Admitted patients occupy a bed for a random amount of time, the length of stay (LOS). We abbreviate the overall average length of stay by ALOS.
To determine appropriate assumptions regarding the arrival process of patients and the LOS distribution, we considered 24 clinical wards of a university medical center, see [4] for details. Based on these data, we make the following assumptions:

Arrivals. It has been widely accepted that unscheduled arrivals occur roughly according to a Poisson process, see e.g. [31]. From [4] it follows that the number of unscheduled

patient admissions per day can also be well approximated by a Poisson distribution in case we distinguish between weekdays and weekends. Motivated by this observation, we here assume that patients arrive according to a time-dependent Poisson process. Besides the weekend effect, the arrival rate may also depend on the day of the week, or the time of day.

LOS. As noted in [4] the ALOS ranges from 1.5 to 7.8 days for the 24 clinical wards[1]. More insights into the LOS characteristics and its relation to the required bed capacity can be obtained by Lorenz curves and Gini coefficients, see [4] or Subsection 2.2 below. Note that there is a discrepancy between the actual (observed) LOS and the medical required LOS. There are several reasons for this discrepancy, including congestion at the subsequent ward (chain effects), no discharges during the night, and less discharges during the weekend.

We use the fact that the LOS distribution can be approximated by a two-phase hyperexponential distribution (H_2). Intuitively, the H_2 assumption indicates that patients can be roughly divided into two different groups, each having an exponential LOS. These groups may for instance represent patients with a short (or normal) LOS and patients with prolonged hospital stay, see also [18, 27]. We refer to Subsection 2.2 for details on the choice of parameters for the H_2 LOS distribution function.

Beds. The capacity of a ward is given by the number of operational beds. In our case, bed capacity is limited by either the available nursing staff or physical constraints. We refer to Section 6 for a brief discussion on nurse staffing. The number of staffed beds fluctuates in practice (illness, deliberately closing beds during the weekends), thus a more flexible setting in which the number of beds may (to some extent) vary over time may be advantageous, see Subsection 5.5. Unless explicitly stated, we however assume that the capacity is fixed at s.

Motivated by the observations above, we use an $M_t/H_2/s/s$ queue to describe the number of patients at a clinical ward. A formal description of this queueing model and an approximation of its performance can be found in Section 4. Next, in Subsection 2.1, we describe a basic clinical ward that is used as an example throughout the paper and detail on the H_2 distribution for the LOS in Subsection 2.2.

2.1 Basic scenario for clinical wards

To demonstrate the effects of time-dependent arrival patterns, we use a basic scenario that represents a typical clinical ward at the university medical center. In particular, we assume that the ALOS is 4 days, which corresponds to the overall average length of stay of the 24 wards that we analyzed. Moreover, we take an average number of arrivals of 6 per day, yielding an average load of $6 \times 4 = 24$. This means that when all patients would be admitted, *on average* 24 beds would be required. A common mistake is to size clinical wards on this average number. This results in operational problems, such as the occurrence of refused admissions.

In Section 5, we consider various modifications of this basic clinical ward. To make good comparisons, we keep the average offered load fixed at 24 in all examples. Unless explicitly stated, we also assume that the number of operational beds is 28. Using the Erlang loss

[1]These numbers are based on data of 2006, but the figures for 2004 and 2005 were analyzed and are comparable.

model, the proportion of refused admissions is 6.7% (e.g., use Equation (4) below) and the bed occupancy is $(6 \times (1 - 0.067) \times 4)/28 \times 100\% \approx 80\%$.
We note that these figures also roughly describe the situation at the Intensive Care Unit of the corresponding hospital.

2.2 LOS and hyperexponential distributions

A good approximation for the LOS is a two-phase hyperexponential distribution function (see also [18, 27] for the occurrence of hyperexponential LOS in health care). Let p_i be the probability that the patient is of type i and let $1/\mu_i$ be the average length of stay for type i, $i = 1, 2$. Since $p_1 + p_2 = 1$, the H_2 distribution is characterized by three parameters. Often, one parameter is eliminated by assuming that $p_1/\mu_1 = p_2/\mu_2$ (balanced means); i.e., the total capacity for both type of patients is equal. Below, we describe a two-parameter fit based on the Gini coefficient (G) and a case where a three-parameter fit is required.

Two-parameter fit.
As mentioned, a two parameter fit is often used in combination with balanced means, i.e., $\mu_1 = \mathbb{E}[S]/(2p_1)$, where $\mathbb{E}[S]$ is the ALOS. In the literature, the parameters are usually determined based on the first two moments (or squared coefficient of variation). An elegant alternative is to use Gini coefficients. The Gini coefficient, traditionally used as a measure of the inequality in wealth, is here used as a measure of dispersion for the inequality in bed capacity. Specifically, a relatively large Gini coefficient (i.e., close to 1) indicates that a disproportional part of the bed capacity is used by a small fraction of the patients, whereas a Gini coefficient of 0 implies that all patients exactly use the same amount of bed capacity. The Gini coefficient thus provides insight into the bed capacity required for patients with prolonged hospital stay. We refer to [4] for a formal definition.
The Gini coefficient for an H_2 distribution does not have a very tractable form in general, but under the assumption of balanced means we have $G = 0.75 - p_1 p_2$, see Appendix A. Hence, $0.5 \leq G < 0.75$ (for $p_1, p_2 > 0$). Using $p_2 = 1 - p_1$, we obtain

$$p_1 = \frac{1}{2} + \sqrt{G - \frac{1}{2}}.$$

The other parameters follow straightforwardly from the assumption of balanced means.
For our university medical center, 22 out of the 24 considered clinical wards have a Gini coefficient between 0.5 and 0.75. For most wards (19 out of the 22), the two-parameter fit gives a good approximation for the LOS distribution. An alternative for the remaining wards, based on a three-parameter fit, is given below.

Three-parameter fit.
In case patients with prolonged stay require a very large part of the bed capacity, which is not uncommon in a university medical center, the assumption of balanced means may not always be appropriate. A prominent example is the ICU. Roughly, we may state that the two-parameter fit is not always satisfactory when the Gini coefficient grows, say, to around 0.7.
An elegant alternative is given in [29], where the parameters are based on the ALOS, squared coefficient of variation, and the proportion of the total mean in the component

with the smaller mean, defined as

$$r = \frac{p_1/\mu_1}{p_1/\mu_1 + p_2/\mu_2}.$$

Note that $r = 0.5$ corresponds to balanced means. In case of disproportional capacity requirements for prolonged hospital stay, r will typically be smaller than 0.5.

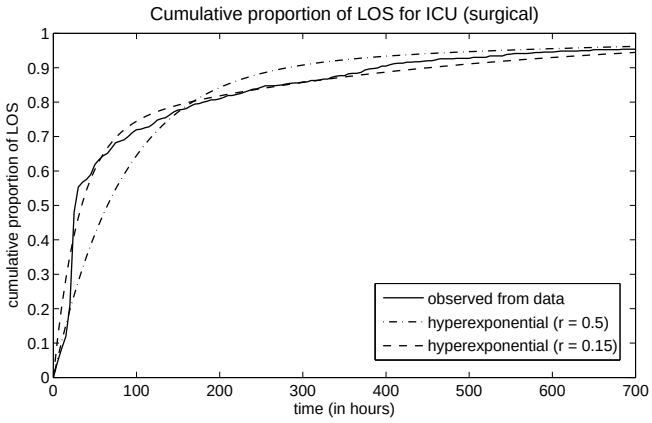

Figure 1: H_2 approximations for the LOS at the ICU surgical.

In Figure 1, the LOS approximation based on $r = 0.5$ and $r = 0.15$ are depicted, along with the observed cumulative LOS for the ICU surgical based on data of 2005. From the figure, it can be immediately concluded that $r = 0.15$ gives a large improvement and a satisfactory LOS approximation.

3 Preliminaries and queueing approximations

Exact results for queueing models with time-dependent arrival patterns are generally difficult to obtain. In this section we define approximations for the $M_t/G/s/s$ models, which are employed in Section 4 for the special case of an $M_t/H/s/s$ model with a piecewise constant arrival rate. In particular, in this subsection, we first review the infinite-server queue $M_t/G/\infty$ with a time-dependent arrival rate, since for that case exact results exist. Second, based on the infinite-server queue, we consider a convenient approximation for the loss fraction (i.e., fraction of refused admissions) in a corresponding loss model. Finally, we discuss some intuitively appealing results on setting staffing levels in service systems. Before presenting the preliminary results, we first introduce some notation. We assume that customers (patients) arrive according to a time-dependent Poisson process with rate $\lambda(t)$ at time t. Let S denote a generic service time (LOS) with mean $1/\mu$. The number of servers (beds) is denoted by s (the infinite-server queue then corresponds to $s = \infty$).

Infinite-server queues
To determine the time-dependent offered load and loss fraction, we rely on approximations based on the infinite-server queue. Using infinite-server queues, we directly obtain the offered load in case there is sufficient bed capacity. Moreover, various performance measures in such $M_t/G/\infty$ queues have a tractable form. For instance, the number of customers

in the $M_t/G/\infty$ queue has a Poisson distribution function with a time-dependent mean $m(t)$; i.e., see [6] and references therein,

$$\mathbb{P}(Q^{\infty}(t) = k) = \frac{m(t)^k}{k!} e^{-m(t)}, \tag{1}$$

where $Q^s(t)$ denotes the number of customers (patients) at time t when the number of beds is s ($s = \infty$ in this case). The mean of the number of customers (patients) in the system at time t can be expressed as

$$m(t) = \mathbb{E}\left[\int_{t-S}^{t} \lambda(u)\mathrm{d}u\right] = \mathbb{E}\left[\lambda(t - S_e)\right]\mathbb{E}[S], \tag{2}$$

where S_e denotes the random variable with the stationary excess distribution of S, i.e.,

$$\mathbb{P}(S_e \leq t) = \frac{1}{\mathbb{E}[S]} \int_0^t \mathbb{P}(S > u)\mathrm{d}u.$$

Using this definition, (2) can be directly rewritten as

$$m(t) = \int_{v=0}^{\infty} \lambda(t - v)\mathbb{P}(S > v)\mathrm{d}v. \tag{3}$$

In case of a cyclic arrival rate, we let a denote the length of a cycle. Also, define $\bar{\lambda} = (1/a)\int_0^a \lambda(t)\mathrm{d}t$ as the average arrival rate, $\bar{\rho} = \bar{\lambda}\mathbb{E}[S]$ as the average instantaneous load, and $\bar{m}$ as the average offered load. It follows from (3) and interchanging integrals that, for cyclic arrival rates,

$$\begin{aligned}
\bar{m} &= \frac{1}{a}\int_0^a m(t)\mathrm{d}t \\
&= \frac{1}{a}\int_0^a \int_0^{\infty} \lambda(t - v)\mathbb{P}(S > v)\mathrm{d}v\mathrm{d}t \\
&= \int_0^{\infty} \mathbb{P}(S > v)\frac{1}{a}\int_0^a \lambda(t - v)\mathrm{d}t\mathrm{d}v \\
&= \int_0^{\infty} \mathbb{P}(S > v)\bar{\lambda}\mathrm{d}v = \bar{\rho}.
\end{aligned}$$

The expression (3) for the time-dependent mean has an even more tractable form in various special cases. The case in which $\lambda(t)$ is a sine function has been analyzed in, e.g., [7]. A disadvantage of a sine (or cosine) function is that cyclic periods that do not have some kind of symmetric behavior are difficult to analyze. As an example, it is not straightforward to take a weekday-weekend pattern into account with this setup. In Section 4, we therefore assume that $\lambda(t)$ is piecewise constant.

Modified-Offered-Load (MOL) approximation
Based on the infinite-server queue, there are different approximations for the time-dependent loss system $M_t/G/s/s$. In [23], the authors show that the so-called Modified-Offered Load (MOL) approximation performs well, especially when the blocking probability is not too large.

In case the arrival rate is constant, the fraction of refused admissions (also called blocking probability) is given by the well-known Erlang pure loss model

$$B(s,\rho) = \frac{\mathbb{P}(Q^{\infty}(t) = s)}{\mathbb{P}(Q^{\infty}(t) \leq s)} = \frac{\rho^s/s!}{\sum_{k=0}^{s} \rho^k/k!}. \tag{4}$$

With MOL, we use the same stationary loss model, but replace the instantaneous offered load by the "modified" offered load $m(t)$ of the infinite-server model with time-dependent arrivals. Denote the approximate blocking probability at time t by B_t. The MOL approximation then reads, see [23],

$$B_t \approx B(s, m(t)) = \frac{m(t)^s/s!}{\sum_{k=0}^{s} m(t)^k/k!}. \tag{5}$$

Number of hospital beds

Determining the number of operational beds is closely related to staffing issues in service systems, such as call centers. In the latter case, the staffing level is in terms of the number of required agents. In particular, for large call centers intuitively appealing results have appeared on optimal staffing levels, see [1, 30].

Let us assume for now that the arrival rate is fixed at λ and the offered load is $\rho = \lambda/\mu$. If the offered load and number of agents are large enough, the staffing level can be determined by the so-called square-root staffing rule, $\beta \in \mathbb{R}$,

$$s = \rho + \beta\sqrt{\rho}. \tag{6}$$

For actual staffing, the obtained s can be rounded to the nearest integer. Obviously, at least ρ agents are required to be able to handle all arriving traffic, explaining the first term on the right-hand-side of (6). The second term $\beta\sqrt{\rho}$ gives the safety staffing, where the parameter β determines the service level. From this formula, the economies of scale become obvious; as the offered load grows, the safety staffing only grows by the square root of the offered load (times β). We refer to [1] for a more elaborate discussion. In Subsection 5.5, we present a procedure based on the square-root staffing rule to easily determine appropriate staffing levels in case of time-dependent arrivals.

First, we note that the square-root staffing rule is based on an asymptotic result as both the offered load and the number of agents (servers) go to infinity [16]. However, this staffing rule also performs well in case the offered load (and number of servers) are not so large. Second, we note that the probability of delay can be held more or less constant in the $M/M/s$ queue for different (large) values of ρ under square-root staffing, see e.g. [1]. Although this appealing property is lost for the $M/G/s/s$ loss model, the scaling in (6) is still of great interest, see e.g. [21]. As already indicated in [2], the blocking probability can then be approximated by

$$B(s,\rho) \approx \frac{\phi(\beta)}{\Phi(\beta)\sqrt{\rho}}, \tag{7}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and density of the standard Normal distribution.

Given an offered load ρ and desired blocking probability α, the approximation in (7) gives an easy way to determine an appropriate β, i.e., the β that satisfies (7). In case the

offered load is small, the approximation (7) is not very accurate. The staffing level can then obviously still be determined by

$$s^* = \arg\min\{s \in \mathbb{N}_+ : B(s, \rho) \leq \alpha\}, \tag{8}$$

with $B(s, \rho)$ given by (4). Routines to find the optimal s^* are not difficult to implement.

4 Time-dependent approximations for clinical wards

In this section, we analyze a queueing model to obtain insights into the effects of time-dependent patterns of arriving patients. The model considered is a special case of the one in Section 3. More specifically, we consider approximations for the $M_t/H/s/s$ queue. That is, patients arrive according to a time-dependent Poisson process with rate $\lambda(t)$. The length of stay has a hyperexponential distribution function consisting of K exponentials. Let $1/\mu$ be the average length of stay and denote the probability of having an exponential LOS with rate μ_i by p_i, $\sum_{i=1}^{K} p_i = 1$. The arriving amount of traffic at time t is given by $\rho(t) = \lambda(t)/\mu$. Let $\bar{\lambda}$ be the average arrival rate and denote the average offered load by $\bar{\rho} = \bar{\lambda}/\mu$. Recall that the number of operational beds is denoted by s and we assume that when a patient arrives and all beds are occupied the arriving patient is refused (e.g., placed at another clinical ward or moved to a different hospital). Note that we assume that once a patient is refused it has no future effect on the considered ward.
To analyze the effects of time-dependent arrivals in clinical wards, we use a more specific arrival rate function. Naturally, we need the arrival rate to be periodic. For both a flexible and tractable setup, we also assume that the arrival rate function is piecewise constant. Suppose that one periodic cycle can be divided into n intervals where the arrival rate is constant. Denote the cycle length by a and let $0 = a_0 < a_1 < \cdots < a_n = a$ be the switching points. For convenience we assume that a periodic interval starts at time 0. Then, $\lambda(t) = \lambda_i$ if $t \in [ka + a_{i-1}, ka + a_i)$ for some integer k and $i \in \{1, \ldots, n\}$.

Remark 4.1. The choice of a piecewise constant arrival rate is convenient for analyzing the effects in different clinical wards. For instance, there can be specific patterns of arriving patients over the day, but also across the week. Studies [3] indicate that a day could be divided in three intervals where the arrival rate is more or less constant, in which case $n = 3$. Depending on the LOS, it might be important to model the number of patients on an hourly basis. This yields $n = 24$ and $a_i = i$ if the time unit is hours. The latter can be important for Emergency Departments where the ALOS is measured in hours.
Another possibility is to let the arrival rate of patients depend on the day within a week, yielding $n = 7$. In case the average number of admitted patients is constant during working days and weekends, respectively, then a convenient choice is $n = 2$, $a_1 = 5$, and $a_2 = 7$, if the time unit is days. ◇

Below, we first analyze the offered load in an $M_t/M/\infty$ model indexed by a parameter $k \in \{1, \ldots, K\}$ representing K different service rates, i.e., we assume the LOS to be exponentially distributed with parameter μ_k. Let $t \in [a_{i-1}, a_i)$ for some $i \in \{1, \ldots, n\}$. For model k, in which case the ALOS is $1/\mu_k$, we denote the load of the system at time t by $m_k(t)$. Using (3), the exponential LOS, and conveniently splitting the integral, we

obtain

$$\begin{aligned} m_k(t) &= \int_{v=0}^{t-a_{i-1}} \lambda_i e^{-\mu_k v} \mathrm{d}v + \int_{v=t-a_{i-1}}^{\infty} \lambda(t-v) e^{-\mu_k v} \mathrm{d}v \\ &= \frac{\lambda_i}{\mu_k} \left(1 - e^{-\mu_k (t-a_{i-1})}\right) + e^{-\mu_k (t-a_{i-1})} \int_{u=0}^{\infty} \lambda(a_{i-1}-u) e^{-\mu_k u} \mathrm{d}u \\ &= \frac{\lambda_i}{\mu_k} \left(1 - e^{-\mu_k (t-a_{i-1})}\right) + e^{-\mu_k (t-a_{i-1})} m_k(a_{i-1}), \end{aligned} \tag{9}$$

where the second step follows from a change of variable and the third step follows from (3).
Using (9), $m_k(a_i)$ can be recursively expressed in $m_k(0)$. In particular, applying (9) for $t = a_1, \ldots, a_i$ and some rewriting yields

$$m_k(a_i) = \sum_{j=1}^{i} \frac{\lambda_j}{\mu_k} e^{-\mu_k (a_i - a_j)} \left(1 - e^{-\mu_k (a_j - a_{j-1})}\right) + e^{-\mu_k a_i} m_k(0).$$

Assuming that the system is in a dynamic steady-state, as discussed in [20] and references therein, we have $m_k(a_n) = m_k(0)$. Hence, taking $i = n$ and some rewriting directly provides the remaining unknown

$$m_k(0) = \frac{\sum_{j=1}^{n} \frac{\lambda_j}{\mu_k} e^{-\mu_k (a - a_j)} \left(1 - e^{-\mu_k (a_j - a_{j-1})}\right)}{1 - e^{-\mu_k a}}. \tag{10}$$

The offered load at time $t \in [a_{i-1}, a_i)$, for some $i \in \{1, \ldots, n\}$, is thus given by

$$m_k(t) = \frac{\lambda_i}{\mu_k} \left(1 - e^{-\mu_k (t-a_{i-1})}\right) + \sum_{j=1}^{i-1} \frac{\lambda_j}{\mu_k} e^{-\mu_k (t-a_j)} \left(1 - e^{-\mu_k (a_j - a_{j-1})}\right) + e^{-\mu_k t} m_k(0),$$

where $m_k(0)$ is given by (10).
The system now considered is an $M_t/H_K/\infty$ system, where the LOS is hyperexponentially distributed with parameters μ_k, p_k, $k \in \{1, \ldots, K\}$. The extension to LOS following a hyperexponential distribution is straightforward. In particular, using the definition (3), we directly obtain

$$m(t) = \int_{v=0}^{\infty} \lambda(t-v) \left(\sum_{k=1}^{K} p_k e^{-\mu_k v} \right) \mathrm{d}v = \sum_{k=1}^{K} p_k m_k(t). \tag{11}$$

This concludes the derivation of the time-dependent offered load.
Now, the MOL approximation for the fraction of refused admissions at time t (B_t) for the $M_t/H_K/s/s$ system is directly obtained by substituting $m(t)$, as given in (11), in (5). To determine the (long-run) average blocking probability $\bar{B}$ we have to take the average number of arrivals during each interval into account. Since the blocking probabilities are statistically identical during each cycle, this leads to the following weighted average:

$$\bar{B} = \frac{1}{\bar{\lambda} a} \int_0^a \lambda(t) B_t \mathrm{d}t.$$

5 Impact of varying arrival rates

In clinical wards the number of arriving patients strongly varies over time. We use a time-dependent Poisson process to model this arrival pattern, where the arrival rate is assumed to be constant on specific intervals. First, the queueing approximation of Section 4 is exploited to analyze the consequences for the offered load in case of no refused admissions. Then we use the queueing model to approximate the long-run fraction of refused admissions and the number of required hospital beds. The results in this section are thus based on the approximation. The approximations are supplemented by simulation results, which are presented in Appendix B.

The arrival pattern typically varies on different time scales. In Subsection 5.1, we consider the daily variation of the arrival process. The behavior across the week is analyzed in Subsection 5.2 and the impact of different operating days is addressed in Subsection 5.3. For convenience, we assume in Subsections 5.1–5.3 that the LOS distribution is exponential (in this case we only require the ALOS as a parameter). In Subsection 5.4, we analyze the impact of different (hyperexponential) LOS distributions. We demonstrate a method to determine the required number of beds across the week in Subsection 5.5. Finally, in Subsection 5.6, we present a counterintuitive example for the theoretical case of deterministic LOS.

5.1 Variation across the day

In this subsection, we consider the effects of the daily variation in arrivals of patients. This daily pattern has been analyzed in e.g. [3, 14]. We consider an $M_t/M/s/s$ queue and assume, for convenience, that there are only two different arrival rates over the day; during office hours (between 8:00 and 18:00 hours) the arrival rate is λ_1, while the arrival rate is λ_2 otherwise (that is, between 18:00 and 8:00 hours).

Based on data of the ICU, we assume that two out of three patients arrive during office hours, thus $\lambda_1 = 14/5 \times \lambda_2$. To demonstrate the effect of the ALOS, we consider three different scenarios that may typically occur in a hospital setting. For a fair comparison, we use the same traffic load $\rho(t)$ for the three scenarios and assume that the average load is 24 as in the basic scenario. For the different scenarios, we take an ALOS of 4 and 1.5 days, and 1.5 hours, respectively. An ALOS of 4 days corresponds to the overall ALOS, while 1.5 days is the minimum ALOS for clinical wards in the hospital under consideration. The fourth scenario is included to demonstrate the time-dependent effects that may occur at an Emergency Department (ED), where treatment times are generally measured in hours, indicating the shorter LOS. The resulting load and approximated blocking probabilities (for $s = 28$) are depicted in Figure 2. We note that the offered load (and thus the average bed occupancy) is most relevant for clinical wards at this time scale. The fraction of refused admissions at this small scale may also be affected by the time of discharges across the day and nurse staffing, see also Remark 5.1 below.

Figure 2 clearly shows that the predictable daily fluctuation in arrival rate of patients have a limited impact on clinical wards where the ALOS is about 4 days (or more). As the ALOS decreases, the impact of variations across the day increases. Intuitively it follows that if the ALOS is in the order of several days, which is the case at most clinical wards, the daily variation in patient arrivals is averaged out. The case in which patients remain in the hospital in the order of several hours is the most difficult case, since this requires

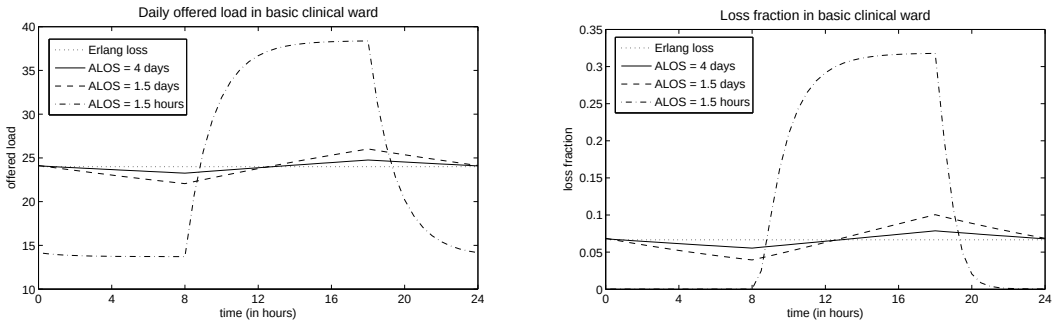

Figure 2: The offered load and fraction of refused admissions (for $s = 28$) across the day.

a full time-dependent analysis. As stated before, this has serious consequences for EDs where fluctuations in workload are common.
The simulation results for the fraction of refused admissions show a similar pattern, see Subsection B.1. The main difference between the simulation and the approximation is that in the simulation the fraction of refused admission faster adapts to the new arrival rate (this may become particularly clear from the ALOS of 1.5 hours). This is due the fact that the MOL approximation is based on the infinite-server queue; the time to reach steady state is generally longer for infinite capacity systems than in case of finite capacity.

Remark 5.1. Although the impact of the daily pattern is limited for the number of required beds across the day, this does not imply that the number of nurses should be held fixed. The staffing level of nurses is often determined using nurse-to-patient ratios, indicating the number of patients a nurse can take care of per shift. This ratio generally differs across the day. For instance, nurse-to-patient ratios during the night shift are often twice as large as the ratios during the early shift, see e.g. [8]. ◇

Time-span and separation of time scales
To theoretically support these observations, consider the $M_t/M/\infty$ queue described in Section 4 with $n = 2$. Clearly, the load attains its minimum and maximum values at $m(0)$ and $m(a_1)$, depending on which λ_i, $i = 1, 2$, is larger. For convenience, assume that $\lambda_1 > \lambda_2$, in which case the load attains its minimum at $m(0)$ and its maximum at $m(a_1)$. Below, we analyze the difference between these loads $m_{\text{span}} = m(a_1) - m(0)$.
Taking $t = a_1$ in (9), $m(a_1)$ can be directly expressed in terms of $m(0)$. Subtracting $m(0)$, using (10) for the case $n = 2$, and some rewriting, provides

$$\begin{aligned} m_{\text{span}} &= \left(1 - e^{-\mu a_1}\right)\left(\frac{\lambda_1}{\mu} - m(0)\right) \\ &= \frac{\lambda_1 - \lambda_2}{\mu}\frac{\left(1 - e^{-\mu a_1}\right)\left(1 - e^{-\mu(a-a_1)}\right)}{1 - e^{-\mu a}}. \end{aligned}$$

Define

$$T_{a_1,a}(\mu) = \frac{\left(1 - e^{-\mu a_1}\right)\left(1 - e^{-\mu(a-a_1)}\right)}{1 - e^{-\mu a}}. \tag{12}$$

The difference between maximum and minimum load can be expressed as

$$m_{\text{span}} = (\rho_1 - \rho_2)\, T_{a_1,a}(\mu). \tag{13}$$

The first term of m_{span} only relates to the difference in load between the two periods. The second term, $T_{a_1,a}(\mu)$, provides insight into the contribution of different time scales. To see this, observe that $T_{a_1,a}(\mu)$ only depends on a_1, a, and the traffic load through μa_1 and μa, that is, through the expected number of service times in the first and total period length. For instance, when $\mu a \to 0$ (and then also $\mu a_1 \to 0$) then the difference in load can be neglected and the model can be approximated by the Simple Stationary Approximation. The explanation for this is quite clear; as the average length of stay is relatively long, the structural variations in the arrival rate occur at a much shorter time scale and average out. A similar result was derived in [7, 11] for a sinusoidal arrival pattern.
In case we fix the period length a and the ALOS $1/\mu$, then it may be readily verified that $T_{a_1,a}(\mu)$ is largest for $a_1 = a/2$, i.e., if the period lengths are equal. In that case, (12) reduces to

$$T_{a/2,a}(\mu) = \frac{1 - e^{-\mu a/2}}{1 + e^{-\mu a/2}}, \tag{14}$$

which only depends on the value of $\mu a/2$.
The simple expressions (12) and (14) quantify in which cases the structural variation in arrival rates may be neglected compared to the differences in load. For this example, with exponential LOS, using the upper bound (14), we have

$$m_{\text{span}} \leq 0.05\,(\rho_1 - \rho_2) \qquad \text{for } \ \mu a \leq 0.199.$$

Thus, as a rule of thumb, we have that in case the ALOS is at least five times as large as the length of the total period, the effect of variations during the cycle are relatively small compared to the differences in instantaneously offered loads.

5.2 Variation across the week

The results of Subsection 5.1 demonstrate that the structural fluctuations in daily admission of patients have a limited impact on the occupancy of most clinical wards. As discussed, this follows from the fact that the ALOS is relatively large compared to the 24-hour cycle. For the predictable variations across the week, the length of the cyclic period is commonly in a similar range as the ALOS. Therefore, this weekly arrival pattern may have a significant impact on the bed occupancy across the week and the staff scheduling. To quantify this impact, we again consider our basic clinical ward. We analyze an $M_t/M/s/s$ queue and assume that during weekdays on average 7.2 patients arrive per day, while during the weekends 3 patients arrive on average per day. The ALOS is again 4 days. These numbers are representative for a clinical ward at the university medical center, see also Subsection 2.1.
From Figure 3 it follows that the load on the system varies significantly over time. In case the number of operational beds is fixed (say, at s) the fraction of refused patients also varies to a large extent. The peak loss fraction based on the approximation is almost 11%. Towards the end of the weekend, the average number of patients at the ward and the loss fraction are relatively small, while these quantities might be undesirably large on Thursdays and Fridays.
The simulation results in Subsection B.2 show an identical pattern. The variation in fraction of refused admissions across the week turns out to be even somewhat larger (e.g., the peak loss fraction then is well over 13%). This can be explained by the finite capacity of a ward, such that the system faster adapts to the changing arrival rate.

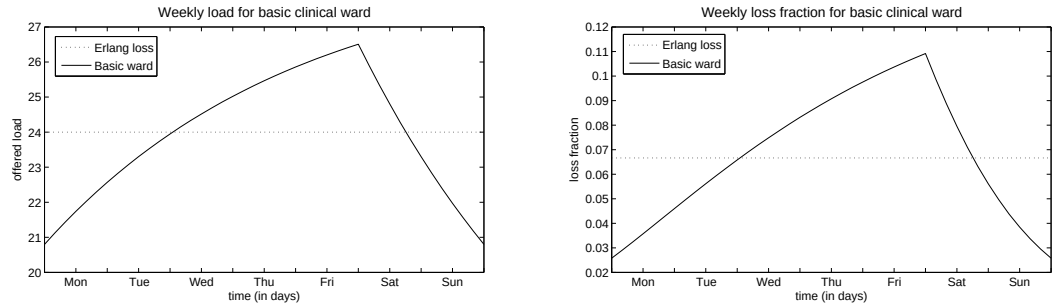

Figure 3: The offered load and the fraction of refused admissions (for $s = 28$) across the week.

The variation in weekly demand also has a negative impact on the total number of refused admissions; the weekly average loss fraction increases from 6.7% to 7.1% when incorporating the non-stationary arrival process. During weekdays, the average loss fraction is 7.4% while it is 5.7% in the weekends. The simulations give a weekly average loss fraction of 7.7%.

Remark 5.2. Ross [26] already conjectured that the loss fraction is larger for a non-stationary arrival process than for a stationary arrival process. This conjecture turns out not to be true in general, see [19] for a counterexample involving deterministic service times. Rolski [25] showed that the conjecture is valid for exponential service times and only one or two servers.
In the Pointwise Stationary Approximation (PSA) the offered load is approximated by the instantaneous traffic load, i.e., $m(t) = \rho(t)$, see e.g. [7, 10, 14]. For the PSA, it is relatively easy to see that the loss fraction increases by incorporating time-dependent arrivals. Fixing s, it has been shown that the throughput $\psi(\rho) = \rho B(s, \rho)$ is a convex function, see e.g. [17, 22]. Since the ALOS is constant, using Jensen's inequality, we have

$$\begin{aligned} B(s, \bar{\rho}) &= \frac{1}{\bar{\rho}} \psi \left(\frac{1}{a} \int_0^a \rho(u) \mathrm{d}u \right) \\ &\leq \frac{1}{\bar{\rho}} \frac{1}{a} \int_0^a \psi(\rho(u)) \mathrm{d}u \\ &= \bar{B}_{\mathrm{PSA}}, \end{aligned}$$

where $\bar{B}_{\mathrm{PSA}}$ is the average loss fraction for the PSA.
Considering the convexity of $\psi(\rho)$, the derivation suggests that the effects of a time-varying arrival rate is largest for values of ρ where $\psi''(\rho)$ is relatively large. Depending on the value of s, this typically happens for moderate to high values of the load (compared to s). In case of our basic clinical ward ($s = 28$), $\psi''(\rho)$ attains its maximum around 22.5. This means that the sharpest increase in loss fraction occurs when the offered load is around 22.5, in which case $B(28, 22.5) \approx 4.5\%$. For clinical wards these are typically the most relevant cases. $\diamond$

To give a more precise description of the bed occupancy across the week, we now consider the daily number of admissions. In Table 1, the proportion of admissions during each day of the week are given for the overall average of the 24 clinical wards and the ICU.
As can be observed from Table 1, the number of admissions slightly varies across weekdays and days in the weekend. For weekdays, this can be explained by the Operating Room

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Overall	18.8 %	16.8 %	16.6 %	16.9 %	14.9 %	7.1 %	9.0 %
ICU	17.1 %	18.6 %	16.7 %	17.5 %	16.0 %	8.2 %	6.0 %

Table 1: Fraction of admissions (in % of total weekly admissions) across the week.

scheduling of different disciplines. Also, note that the average number of admissions on Friday is lower than on the other weekdays. For the average over the 24 wards, more patients are admitted on Sundays than on Saturdays. This difference is likely to be caused by the fact that patients having a scheduled surgery on Monday are sometimes admitted on Sunday. Such patients are usually placed at a normal care facility, explaining the difference between the overall average and the ICU during weekends.
To illustrate the impact of fluctuations in daily admissions, we consider our basic clinical ward where the arrival pattern is as in Table 1 (with an average of 6 arrivals per day); see Figure 4 for the result. Note that the time-dependent offered load and loss fraction are very similar to those in Figure 3, indicating that the weekday-weekend pattern has a predominant effect.

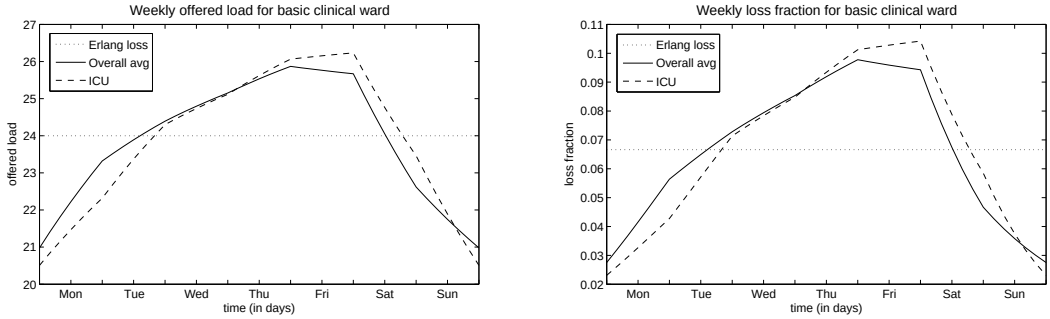

Figure 4: The offered load and the fraction of refused admissions (for $s = 28$) for daily admissions.

When the number of beds is fixed, then the time-dependent admissions have a clear negative impact on the performance of clinical wards. However, the time-dependent pattern can also be exploited for determining appropriate staffing levels. Since a large part of the personnel might want to have days off during the weekend, the number of staffed beds could be reduced during that period. The structural reduction of the number of operational beds during the weekend is already standard procedure at the ICU of the considered university medical center. This option is further exploited in Subsection 5.5.
Although we only presented one numerical example here, we note that this weekly behavior seems typical for many clinical wards, which is also recognized by hospital professionals.

5.3 Operating Room schedule

The Operating Room (OR) schedule also has an effect on the number of arrivals to a clinical ward across the week. In this subsection, we explore the impact of the OR schedule on the occupancy of clinical wards.
In general, each surgical discipline gets one or more rooms assigned to for a fixed number of days (referred to as OR sessions) during the week. In the medical center of this study, the

average number of OR sessions per surgical discipline is 5 per week. Note that this does not necessarily mean that a surgical group gets 1 session assigned per weekday. Surgeons in a university medical center have many obligations such as: teaching medical students, appointments at the outpatient clinic, doing research and operating patients. This directly leads to variability in the number of OR sessions per medical discipline across the week. As patients are usually admitted the day of surgery or the day before, the number of arrivals is directly affected by the OR schedule.
To investigate the impact of the OR schedules, we analyze an $M_t/M/s/s$ queue with three scenarios as indicated in Table 2. The first scenario corresponds to the first scenario of Subsection 5.2 with on average 7.2 and 3 arrivals per day during the week and weekend, respectively, and is included for reference. The second and third scenario represent an OR schedule which is heavy at the beginning and end of the week, respectively. The difference of (on average) two patients on two weekdays, are based on practical experience in our medical center.

Scenario	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week-weekend	7.2	7.2	7.2	7.2	7.2	3.0	3.0
Mon, Tue	8.4	8.4	6.4	6.4	6.4	3.0	3.0
Thu, Fri	6.4	6.4	6.4	8.4	8.4	3.0	3.0

Table 2: Average number of daily admissions.

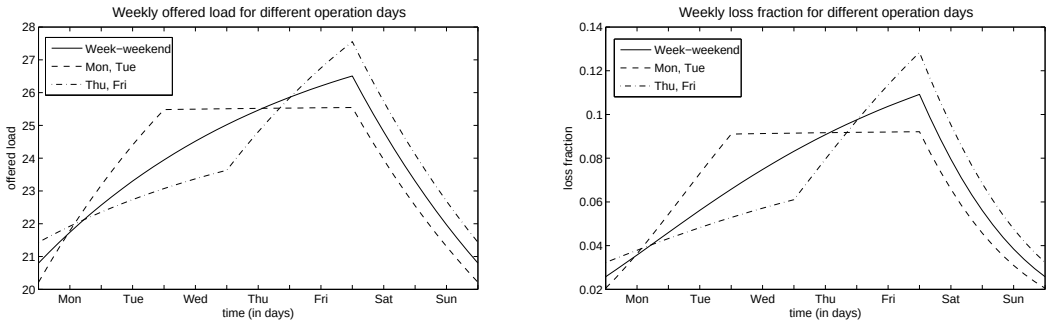

Figure 5: The offered load and the fraction of refused admissions (for $s = 28$) for different operating days.

The resulting offered load and proportion of refused admissions for 28 operational beds can be found in Figure 5. As can be observed from the figure, a busy OR schedule at the beginning of the week (say, on Monday and Tuesday) leads to a more balanced required number of operational beds across the week. In case the OR schedule becomes dense towards the end of the week, there is a sharper peak in the loss fraction (and thus required number of beds) on Thursday and Friday. We note that the simulations show a similar pattern for fraction of refused admissions, see Subsection B.3. Again, the variation in loss fraction is somewhat larger for the simulation experiments than for the approximations.
The above analysis only serves as an example of the impact of the OR schedule on clinical wards for situations that often occur in practice. To determine the optimal OR schedule from the perspective of clinical wards, many factors have to be taken into account, including a thorough flow analysis of all surgical clinical wards. Most important is to be aware

of the impact of the OR schedule on offered load and refused admissions at clinical wards.

5.4 Variability in LOS

In Subsections 5.1-5.3, we assumed for convenience that the LOS follows an exponential distribution. In this subsection we analyze the impact of the LOS distribution, since an approximation based on the hyperexponential distribution is generally better. Specifically, we analyze the $M_t/H/s/s$ queue.
We consider three different cases for the distribution of the LOS; (i) an exponential LOS, (ii) a hyperexponential LOS with $r = 0.5$ and a squared coefficient of variation of 4, and (iii) a hyperexponential LOS with $r = 0.15$ and the same squared coefficient of variation. Cases (i) and (ii) typically correspond to clinical wards where the capacity for patients with prolonged stay is low to moderate and case (iii) is typical for an ICU. We use again our basic clinical ward with the weekday-weekend pattern of Subsection 5.2 as a reference scenario. That is, the arrival rate is 7.2 during weekdays and 3 during weekends. The ALOS is equal to 4 days.
The offered load and proportion of refused admissions (for 28 beds) across the week for the different LOS distributions are depicted in Figure 6. In the left part of the figure, we also included the offered load in case of a deterministic LOS (see Subsection 5.6 for the proportion of refused admissions). The distribution has a clear effect on the time-dependent performance. The differences in load are most extreme for the case of a fixed (deterministic) LOS. Also observe the stabilizing effect of the patients with a prolonged stay on the offered load and refused admissions across the week; the differences in load and loss fraction across the week are smaller for the hyperexponential distribution, than those for the exponential.

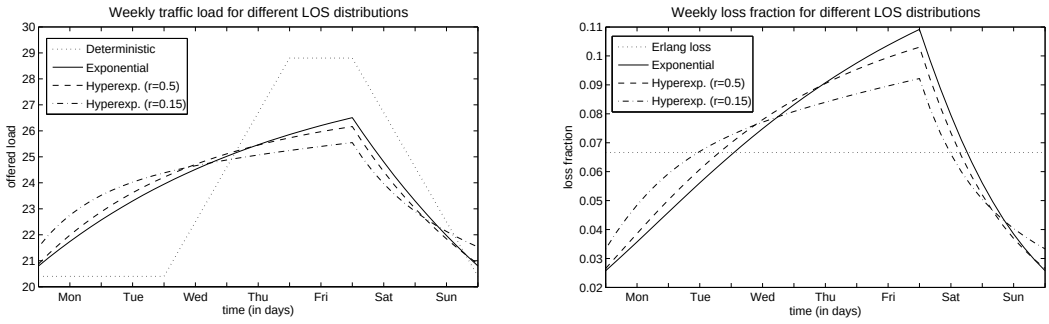

Figure 6: The offered load and the fraction of refused admissions (for $s = 28$) for different LOS distributions.

More generally, it can be observed from Figure 6 that the peak loss fraction increases as the variability in the LOS decreases, i.e., a more variable LOS has a stabilizing effect on the time-dependent bed capacity. The simulation results yield the same conclusion (see Subsection B.4). This follows from the approach to steady state, see [5]. Intuitively, lower variability implies more responsiveness. For example, in case of deterministic service times the offered load is adapting fastest to the changing arrival rate. This observation stems from [5] for the case of a sinusoidal arrival rate function.
For the average loss fraction there is also a slight difference between the simulations and the MOL approximation. For MOL, the average loss fraction is around 7.1% for the

exponential and hyperexponential distributions. For the simulations, the average loss fraction roughly equals 7.7%, 7.7%, and 7.3% for cases (i), (ii), and (iii), respectively. Recall that the loss fraction for the stationary model is roughly 6.7%.
From the analysis above, it can be concluded that reducing the variability in LOS increases the variability in bed occupancy. It should be noted though, that a reduction in the LOS variability is in practice often accompanied with a reduction in the average LOS. Significant gains in terms of number of operational beds can be obtained by such a reduction in the ALOS.

Remark 5.3. The difference between the maximum and minimum offered load can also easily be analyzed for the hyperexponential distribution. In particular, for the case $n = 2$ and an H_K LOS distribution, it follows from (11) and a similar analysis as for (13) that

$$\begin{aligned} m_{\text{span}} &= \sum_{k=1}^{K} p_k \left(\frac{\lambda_1}{\mu_k} - \frac{\lambda_2}{\mu_k} \right) T_{a_1,a}(\mu_k) \\ &= (\lambda_1 - \lambda_2) \sum_{k=1}^{K} \frac{p_k}{\mu_k} T_{a_1,a}(\mu_k), \end{aligned}$$

with $T_{a_1,a}(\mu_k)$ given by (12).
Fixing a_1, a, it follows from $\sum_{k=1}^{K} p_k/\mu_k = 1/\mu$ and the convexity of $T_{a_1,a}(\cdot)$ that

$$(\lambda_1 - \lambda_2) \sum_{k=1}^{K} \frac{p_k}{\mu_k} T_{a_1,a}(\mu_k) \;=\; (\lambda_1 - \lambda_2) \frac{1}{\mu} \sum_{k=1}^{K} \frac{p_k}{\mu_k} \mu T_{a_1,a}(\mu_k) \;\leq\; (\lambda_1 - \lambda_2) \frac{1}{\mu} T_{a_1,a}(\mu).$$

Hence, the variability in the offered load is smaller for a hyperexponential LOS than for an exponential LOS, as we also observed in Figure 6. ◇

5.5 Number of operational beds

As shown in Subsection 5.2, the bed occupancy severely fluctuates in case the average number of daily admissions varies across the week. If the number of staffed beds is kept fixed, the fluctuations in bed occupancy can have an undesirable effect on the number of refused admissions during congested periods. This effect can easily be resolved by allowing for a more flexible staffing across the week. The analysis is based on an $M_t/H/s_t/s_t$ model. Generally, the number of required beds can easily be determined by an application of the Erlang loss formula. Another easy and insightful rule for determining the number of staffed beds is the square-root staffing formula, see Section 3. The square-root staffing approximation performs well if the number of beds is not too small and can therefore be used for reasonably sized wards. Moreover, the square-root staffing formula can be easily applied when the average bed occupancy is non-constant (as in the case of time-dependent arrivals).
As mentioned in Section 3, we first need to determine the parameter β corresponding to the offered service level. The average number of available beds is usually a strategic decision while the structural opening and/or closing of beds during the week is of a tactical nature. Following this hierarchy the operational planning involves day-to-day decisions on the number of beds. In particular, the first issue is typically a managerial trade-off between service level and efficiency. The Erlang loss formula can often very well be applied to

assign numerical values to this trade-off, see [4, 24]. This trade-off also typically implies a decision on β; rewriting of (6) yields

$$\beta = \frac{s^* - \rho}{\sqrt{\rho}},$$

where s^* is the average number of beds decided upon.
The required number of beds across the week can now be directly determined by (6), i.e., $s(t) = m(t) + \beta\sqrt{m(t)}$, with β given as above and $s(t)$ the number of beds at time t. For the basic ward ($s^* = 28$ and $\bar{\rho} = 24$), we get $\beta \approx 0.81$. In Figure 7, we consider the week-weekend pattern as in Subsection 5.2, i.e., an average of 7.2 and 3 arrivals per day during the week and weekends, respectively. In particular, we depicted the number of beds as prescribed by the square-root staffing formula and the corresponding loss fraction for two LOS distributions: (i) an exponential and (ii) a hyperexponential with $r = 0.15$ (corresponding to an ICU). For both LOS distributions, the average number of beds roughly equals 28. For an exponential LOS, the number of beds varies between 25 and 31. The time-dependent loss fraction lies between 5.8% and 7.6% across the week, with an average of 6.7%. For an hyperexponential distribution, the number of beds varies between 25 and 30 and the numbers for the loss fraction are comparable.

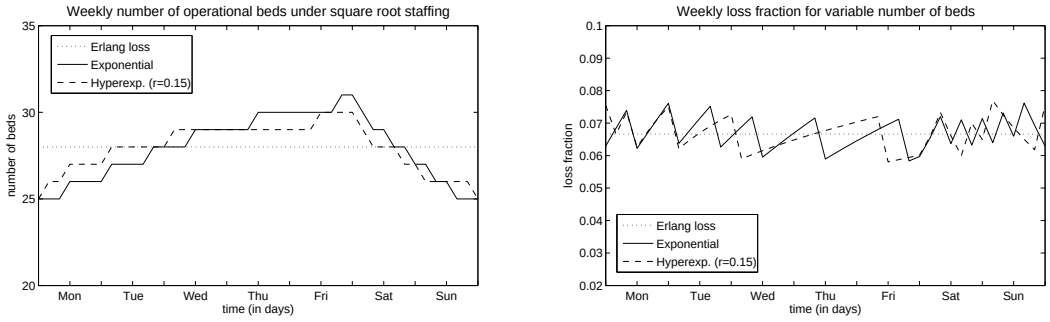

Figure 7: The number of beds across the week based on the square-root staffing.

The simulation results indicate a similar pattern. The loss fraction in the simulation experiments is more responsive to changes in the number of staffed beds than the MOL approximation, yielding a sharper increase (decrease) in the four-hour period of a bed reduction (bed increase). It seem unlikely that such short-term effects will occur in practice with opening or closing of beds.
For practical purposes, it might not always be desirable to change the actual number of open beds multiple times during the week. The prescribed staffing levels should therefore merely serve as a guideline.

Remark 5.4. Some care is required when applying the square-root staffing formula for Erlang loss models. For Erlang delay models, square-root staffing allows for convenient interpretations in terms of keeping the probability of delay fixed. From (7) it easily follows that the loss fraction decreases as ρ (and thus s) increase. When the fluctuations in the offered load are not too large, which is typically the case for week-weekend patterns in clinical wards, the differences in loss fractions remain however small. ◇

Remark 5.5. The range of different offered loads m_{span} (see Remark 5.3) directly provides an indication to what extent the required number of beds may vary. In particular, if the

offered load is not too small and the load variations are not too large (such that the differences in $\sqrt{m(t)}$ are small), the number of staffed beds can also be approximated by $s = m(t) + \beta\sqrt{\rho}$. In that case, m_{span} gives the difference between the maximum and minimum number of required beds. ◇

5.6 Deterministic LOS and the limitation of approximations

In this subsection we consider the $M_t/D/s/s$ queue, i.e., we assume that the LOS is fixed at 4 days. We use again our basic clinical ward of Subsection 5.2 as a reference scenario. That means that the arrival rate is 7.2 during weekdays and 3 during the weekend. We stress that a deterministic LOS is of limited practical relevance, but provides an interesting theoretical example.

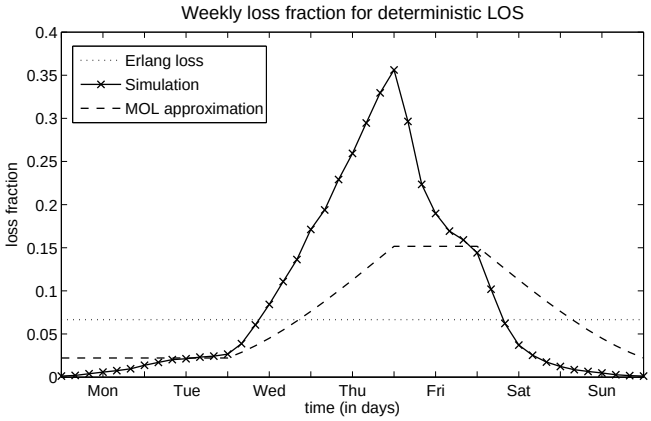

Figure 8: Fraction of refused admissions; MOL and simulation results.

The offered load across the week is depicted at the left-hand side of Figure 6, see Subsection 5.4. The MOL approximation is based on the offered load and is shown in Figure 8. Moreover, the fraction of refused admissions is also determined using simulation, see also Figure 8 for the results. 0 The simulation results and the differences with MOL are remarkable. Most striking is the blocking probability of roughly 35% towards the end of Thursday. This is considerably larger than the blocking probability of 15.2% provided by MOL. We note that for a stationary arrival process at rate 7.2 (also during the weekend) the blocking probability also equals 15.2%. Hence, reducing the arrival rate during the weekend leads to a large increase of the blocking probability on Thursday.

This remarkable result can be explained as follows. The reduction of the number of arrivals during the weekend leads to a relatively empty ward on Monday. As a consequence, hardly any patients are blocked on Monday and Tuesday. Since the LOS is exactly 4 days, all patients that arrive on Monday and Tuesday still occupy their beds on Thursday. In case the arrival rate would have been 7.2 during the weekend, more patients would have been blocked on Monday and Tuesday, thereby emptying beds for patients arriving later during the week.

The simulation gives an average blocking probability of 10.7%. This is indeed considerably larger than the average blocking probability based on MOL, which is 7.3%. Thus, the simulation results strengthen the conclusion of Subsection 5.4 that reducing the variability

in LOS increases the variability in bed occupancy.

6 Conclusion and discussion

In this paper we considered the impact of predictable fluctuations in the average number of arrivals over time for clinical wards. These predictable fluctuations include both daily and weekly patterns. Using queueing approximations based on the infinite-server queue, we showed that the daily pattern, in which most patients arrive during office hours, has a limited impact on the average number of occupied beds at clinical wards. In contrast, the daily fluctuations are crucial for capacity planning at the Emergency Department. As a rule of thumb, we state that the impact of fluctuations is relatively small if the average length of stay exceeds five times the cycle length at which these fluctuations occur. In that case, such fluctuations average out.

From the queueing approximation and simulations, it follows that the weekly pattern, in which most patients arrive during weekdays, has a significant impact on the required bed capacity at clinical wards. The offered load and fraction of refused admissions may become undesirably large towards the end of the week. From the approximation it also follows that an increased variability in LOS has a stabilizing effect on the variability in refused admissions. In the theoretically interesting case of a deterministic LOS, the fraction of refused admissions may even become unexpectedly large on Thursdays. Finally, we indicated how flexible staffing can partly eliminate operational problems such as variability in refused admissions.

We like to stress that the presented numerical results should serve as an *indication* of the impact of the predictable arrival pattern of patients. Depending on the characteristics of a specific clinical ward, the queueing approximation can easily be implemented in an Excel spreadsheet to give useful insight into the weekly variability in the offered load and fraction of refused admissions. They also prescribe the required number of beds to keep the user-defined fraction of refused admissions stable over time. We note that for the practically relevant cases of (hyper)exponentially distributed LOS, the queueing approximations perform reasonably well compared to simulations. Finally, the approximations can also be utilized in, for instance, an optimization setting.

A subsequent step is to determine the required staffing level. This is typically accomplished by using a fixed ratio for the number of patients (beds) that a single nurse can take care off during a shift, i.e., the nurse-to-patient ratio. The nurse-to-patient ratios generally differ between different shifts across the day and between weekdays and weekends. We refer to [8] and references therein for a more elaborate discussion. Two recent studies [15, 28] propose to use queueing models to determine appropriate staffing levels in order to provide timely responses to patients needs. In particular, the study [28] shows that using a fixed ratio has undesirable effects. In this paper, we focus on the number of beds from which appropriate staffing levels can be determined.

Furthermore, some data experiments indicate that the number of discharges also varies across different days in the week. For an accurate modeling of the offered load and number of refused admissions across the week, the discharge policy should be taken into account as well. Therefore, the results in this paper mainly serve as an indication of the impact of weekend effect on the number of beds required and to stress the importance of taking time-dependent variations into account.

A Gini coefficient and H_2 distributions

Let S be a random variable having a two-phase hyperexponential (H_2) distribution function, which corresponds to the approximation of the LOS (see Subsection 2.2). Since the H_2 distribution is piecewise differentiable, the Gini coefficient is given by

$$G = 1 - \frac{1}{\mathbb{E}[S]} \int_0^\infty \mathbb{P}(S > y)^2 \mathrm{d}y.$$

Using the H_2 excess distribution, we have

$$\begin{aligned} G &= 1 - \frac{1}{\mathbb{E}[S]} \int_0^\infty \left(p_1 e^{-\mu_1 y} + p_2 e^{-\mu_2 y}\right)^2 \mathrm{d}y \\ &= 1 - \frac{1}{\mathbb{E}[S]} \left(\frac{p_1^2}{2\mu_1} + \frac{p_2^2}{2\mu_2} + \frac{2p_1p_2}{\mu_1+\mu_2}\right), \end{aligned} \tag{15}$$

with $\mathbb{E}[S] = p_1/\mu_1 + p_2/\mu_2$.
Under the assumption of *balanced means*, i.e., $p_1/\mu_1 = p_2/\mu_2$, the above expression can be simplified. In particular, for the final term in (15), we may write

$$\frac{2p_1p_2}{\mu_1+\mu_2} = \frac{p_1p_2\frac{2p_1}{\mu_1}}{p_1 + \frac{\mu_2}{\mu_1}p_1} = \frac{p_1p_2\mathbb{E}[S]}{p_1+p_2} = p_1p_2\mathbb{E}[S].$$

Using $\mathbb{E}[S] = 2p_1/\mu_1 = 2p_2/\mu_2$ once more, (15) can be rewritten as

$$\begin{aligned} G &= 1 - \frac{1}{\mathbb{E}[S]} \left(\frac{1}{4}\mathbb{E}[S] + p_1p_2\mathbb{E}[S]\right) \\ &= \frac{3}{4} - p_1p_2. \end{aligned}$$

Hence, in case of balanced means and $p_1, p_2 > 0$, we have $G \in [0.5, 0.75)$.

B Simulation results

In this part, we present the fraction of refused admissions for the different scenarios of Section 5. In particular, the scenarios simulated in Section B.1–B.5 correspond to the scenarios used in Subsections 5.1–5.5 for the approximations.

B.1 Variation across the day

In this subsection we consider the same scenario as in Subsection 5.1. That is, we assume that the average load is 24 and the number of beds is 28. Two out of three patients are assumed to arrive during office hours and we consider three different ALOS: 4 days, 1.5 days, and 1.5 hours.
In the simulation, we have divided each day into 24 one-hour periods. We have used a warm-up time of 200 days and a simulation time of 40.000 days. The simulation results in Figure 9 show that the fraction of refused admissions adapts slightly faster to changing arrival rates than the MOL approximation prescribes. This follows from the fact that the approximation is based on the infinite-server queue.

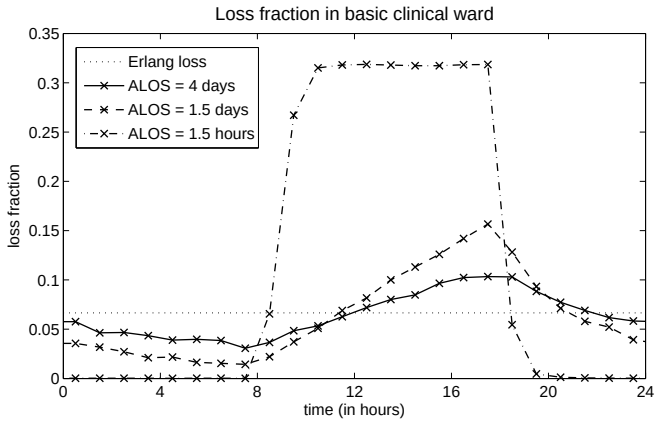

Figure 9: Simulated fraction of refused admissions across the day.

B.2 Variation across the week

This subsection supplements the approximation results of Subsection 5.2. Again, the arrival rate is 7.2 during weekdays and 3 during the weekend. The ALOS is 4 and the number of beds is 28.

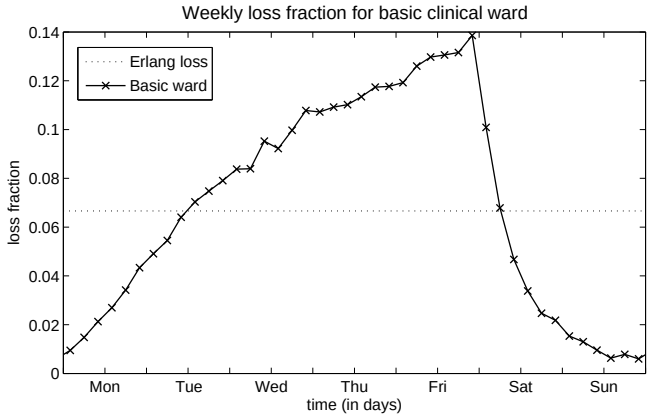

Figure 10: Simulated fraction of refused admissions across the week.

In the simulation, each day is divided into 6 four-hour intervals. We used a warm-up period of 100 weeks and a simulation time of 20.000 weeks. The resulting fraction of refused admissions (see Figure 10) shows an identical pattern as those based on the MOL approximation. The variation in refused admissions in the simulation is slightly larger though, which can again be explained by the fact that MOL is based on the infinite-server queue.

B.3 Operating Room schedule

In this subsection we present the simulation results corresponding to Subsection 5.3. The arrival rates are as in Table 2. The ALOS is 4 days and the number of beds is 28.

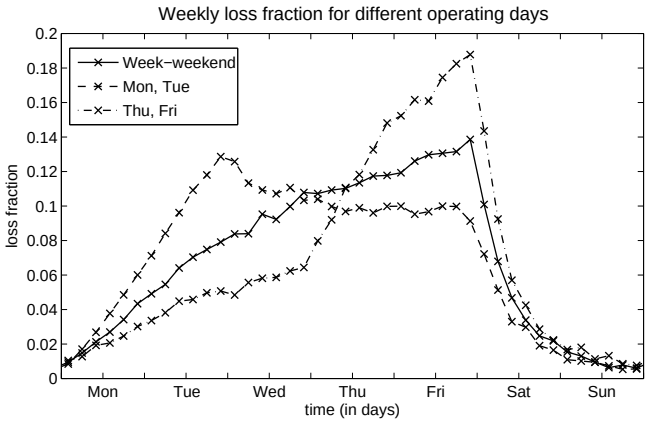

Figure 11: Simulated fraction of refused admissions for different operating days.

As in Subsection B.2, each day is divided into 6 equally spaced intervals. We have simulated the system for 20.000 weeks with a warm-up period of 100 weeks. The loss fraction in each interval for the three different scenarios are depicted in Figure 11 and show a similar pattern as those based on MOL.

B.4 Variability in LOS

In this subsection, we present the difference in loss fraction for exponential and hyper-exponential LOS distribution based on simulation (see Subsection 5.6 for a fixed LOS). Again, the arrival rate is 7.2 and 3 during weekdays and days in the weekend, respectively. The ALOS is 4 days and the number of beds equals 28.

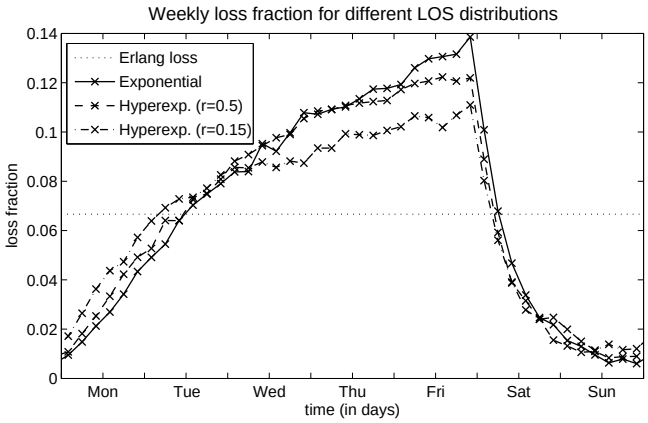

Figure 12: Simulated fraction of refused admissions for different LOS distributions.

We used four-hour periods again, and have simulated 20.000 weeks with a warm-up period of 100 weeks. The fraction of refused admissions in each interval for the different LOS distributions are given in Figure 12. As we concluded based on MOL, an increased variability in LOS results in a reduced variability in fraction of refused admissions across

the week.

B.5 Number of operational beds

This subsection supplements the results of Subsection 5.5. In particular, we consider the week-weekend pattern in which the arrival rate equals 7.2 and 3 during weekdays and days in the weekend, respectively. The ALOS is 4 days and the number of beds is as prescribed in the left part of Figure 7. In the simulation, we assumed that only arriving patients are refused. In case the number of beds should be reduced when all beds are occupied, then the first beds are removed that become available.

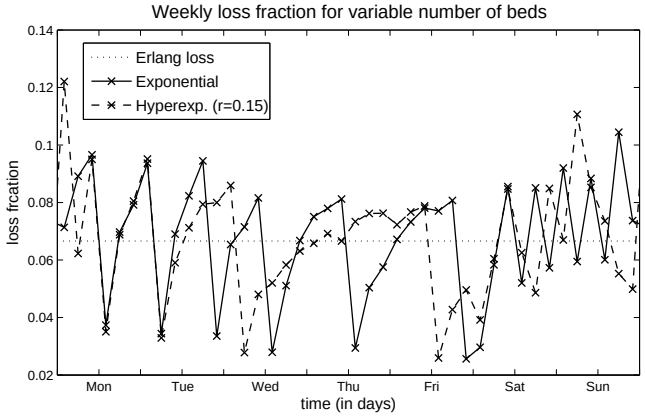

Figure 13: Simulated fraction of refused admissions for variable number of beds.

Using four-hour intervals, we have simulated the system for 20.000 weeks (the warm-up time equals 100 weeks). The fraction of refused admissions is given in Figure 13. Note that in the four-hour period of a bed reduction, the loss fraction increases considerably after which it becomes more stable. Similarly, in the four-hour period of a bed increase, the loss fraction drops significantly, after which it stabilizes. The loss fraction is thus, again, more responsive for the simulation experiments than for the MOL approximation. It is unlikely though that these extremely short-term effects occur in practice with opening or closing of beds.

Acknowledgment

The authors would like to thank two anonymous referees for their valuable comments. The authors are also grateful to MSc. Lillian van Zanten for her support regarding the extensive data analysis and carrying out some early experiments on the impact of time-dependent arrivals for clinical wards and to Prof. G.M. Koole for having stimulating discussions with us. Finally, the authors would like to thank MSc. Paulien Out and MSc. Alex Roubos for their fruitful assistance in carrying out the simulation experiments.

References

[1] Borst, S.C., A. Mandelbaum, M.I. Reiman (2004). Dimensioning large call centers. *Operations Research* **52**, 17–34.

[2] Brockmeyer, E., H.L. Halstrøm, A. Jensen (1948). The life and works of A.K. Erlang. Transactions of the Danish Academy of Technical Sciences **2**, Denmark.

[3] de Bruin, A.M., A.C. van Rossum, M.C. Visser, G.M. Koole (2007). Modeling the emergency cardiac in-patient flow: An application of queueing theory. *Health Care Management Science* **10**, 125–137.

[4] de Bruin, A.M., R. Bekker, L. van Zanten, G.M. Koole (2008). Dimensioning clinical wards using the Erlang loss model. To appear in *Annals of Operations Research*.

[5] Davis, J.L., W.A. Massey, W. Whitt (1995). Sensitivity to the service-time distribution in the non-stationary Erlang loss model. *Management Science* **41**, 1107–1116.

[6] Eick, S.G., W.A. Massey, W. Whitt (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research* **41**, 731–742.

[7] Eick, S.G., W.A. Massey, W. Whitt (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* **39**, 241–252.

[8] Elkhuizen, S.G., G. Bor, M. Smeenk, N.S. Klazinga, P.J.M. Bakker (2007). Capacity management of nursing staff as a vehicle for organizational improvement. *BMC Health Services Research* **7**:196.

[9] Feldman, Z., A. Mandelbaum, W.A. Massey, W. Whitt (2005). Staffing of time-varying queues to achieve time-stable performance.

[10] Green, L.V., P.J. Kolesar (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* **37**, 84–97.

[11] Green, L.V., P.J. Kolesar, A. Svoronos (1991). Some effects of nonstationarity on multiserver Markovian queueing systems. *Operations Research* **39**, 502–511.

[12] Green, L.V., P.J. Kolesar, J. Soares (2001). Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* **49**, 549–564.

[13] Green, L.V., V. Nguyen (2001). Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* **36**, 421–442.

[14] Green, L.V., P.J. Kolesar, W. Whitt (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16**, 13–39.

[15] Green, L.V., N. Yankovic (2008). A queueing model for nurse staffing.

[16] Halfin, S., W. Whitt (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29**, 567–588.

[17] Harel, A. (1990). Convexity properties of the Erlang loss formula. *Operations Research* **38**, 499–505.

[18] Harrison, G.W. (2001). Implications of mixed exponential occupancy distributions and patient flow models for health care planning. *Health Care Management Science* **4**, 37–45.

[19] Heyman, D.P. (1982). On Ross's conjectures about queues with nonstationary Poisson arrivals. *Journal of Applied Probability* **19**, 245–249.

[20] Heyman, D.P., W. Whitt (1984). The asymptotic behavior of queues with time-varying arrival rates. *Journal of Applied Probability* **21**, 143–156.

[21] Jansen, A.J.E.M., J.S.H. van Leeuwaarden, A.P. Zwart (2007). Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Advances in Applied Probability*, to appear.

[22] Krishnan, K.R. (1990). The convexity of loss rate in an Erlang loss system and sojourn in an Erlang delay system with respect to arrival and service rates. *IEEE Transactions on Communications* **38**, 1314–1316.

[23] Massey, W.A., W. Whitt (1994). An analysis of the Modified Offered Load approximation for the Erlang loss model. *Annals of Applied Probability* **4**, 1145–1160.

[24] McManus, M.L., M.C. Long, A. Copper, E. Litvak (2004). Queueing theory accurately models the need for critical care resources. *Anesthesiology* **100**, 1271–1276.

[25] Rolski, T. (1984). Comparison theorems for queues with dependent inter-arrival times. In F. Baccelli, G. Fayolle, editors, *Modelling and performance evaluation methodology*, 42–67. Springer, Berlin.

[26] Ross, S.M. (1978). Average delay in queues with non-stationary Poisson arrivals. *Journal of Applied Probability* **15**, 602–609.

[27] Taylor, G.J., S.I. McClean, P.H. Millard (2000). Stochastic models of geriatric bed occupancy behaviour. *Journal of the Royal Statistical Society A* **163**, 39–48.

[28] de Véricort, F., O.B. Jennings (2008). Nurse-to-patient ratios in hospital staffing: a queueing perspective.

[29] Whitt, W. (1984). On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Technical Journal* **63**, 163–175.

[30] Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science* **38**, 708–723.

[31] Young, J.P. (1965). Stabilization of inpatient bed occupancy through control of admissions. *Journal of the American Hospital Association* **39**, 41–48.