

Scheduling admissions and reducing variability in bed demand

René Bekker · Paulien M. Koeleman

Received: 25 August 2010 / Accepted: 11 May 2011 / Published online: 11 June 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Variability in admissions and lengths of stay inherently leads to variability in bed occupancy. The aim of this paper is to analyse the impact of these sources of variability on the required amount of capacity and to determine admission quota for scheduled admissions to regulate the occupancy pattern. For the impact of variability on the required number of beds, we use a heavy-traffic limit theorem for the $G/G/\infty$ queue yielding an intuitively appealing approximation in case the arrival process is not Poisson. Also, given a structural weekly admission pattern, we apply a time-dependent analysis to determine the mean offered load per day. This time-dependent analysis is combined with a Quadratic Programming model to determine the optimal number of elective admissions per day, such that an average desired daily occupancy is achieved. From the mathematical results, practical scenarios and guidelines are derived that can be used by hospital managers and support the method of quota scheduling. In practice, the results can be implemented by providing admission quota prescribing the target number of admissions for each patient group.

Keywords Variability in bed demand · Admission scheduling · Capacity requirement · Queueing theory · Heavy-traffic approximation · Quadratic programming

R. Bekker (✉)
Department of Mathematics, VU University Amsterdam,
De Boelelaan 1081, The Netherlands
e-mail: rbekker@few.vu.nl

P. M. Koeleman
VU University Amsterdam & CC Zorgadviseurs,
Houttuinlaan 14, Woerden, The Netherlands

1 Introduction

With the growing demand for health care resources, the pressure on the efficient usage of the available bed capacity is increasing. The workload at clinical wards is most often highly variable, leading to the need for extra capacity to respond to peaks in demand for beds. In addition to these extra capacity requirements, the variability in workload does have other negative side effects. For instance, Litvak et al. [16] show that reducing variability in bed demand helps to reduce the stress of the nursing staff and to improve the safety of patients.

Surprisingly, studies have shown that the variation in the number of scheduled patients admitted is generally at least as large as the variation in the number of emergency admissions, and often larger, see e.g. McManus et al. [20] and de Bruin et al. [5]. The variability in admissions leads to highly variable bed occupancy. The admission process is also largely affected by the schedule of the Operating Theater (OT). The OT schedule allocates the available operating time to the different surgical disciplines, but, in most cases, it does not specify which or how many procedures are to be executed in the allocated times and so the number of patients admitted on every weekday can vary significantly. This OT schedule results in a weekly bed occupancy pattern, but the number of occupied beds on each day can still vary significantly from week to week. Moreover, during the weekend the number of elective admissions is generally very small, leading to extra workload fluctuations over the week.

A simple way to reduce the variability would be to admit a fixed number of patients every day of the week. This could be implemented using a fixed quatum

for the number of daily admissions, thereby removing any unnecessary variation in demand. The remaining variation would solely be due to emergency arrivals and variations in the length of stay in the hospital.

The absence of a substantial number of scheduled admissions during the weekend complicates the use of a fixed quatum per day. In addition, it is current practice in most hospitals that the number of staffed beds is lower during weekends, partly because of higher staffing costs. A fixed daily quatum (for every day of the week) would not accommodate this, but yields the same expected bed demand every day of the week. An alternative approach is to use different admission quota for the days of the week, taking differences in length of stay (LOS) between patient types into account. In this paper we determine the number of scheduled admissions for every day of the week, with the objective of keeping the bed demand as close as possible to some target load. This target can be different for every day of the week, thereby accommodating a lower number of staffed beds during the weekends. The result will be quota for the number of scheduled admissions for different patient types on every day of the week.

The main contribution of this paper is to determine admission quota for scheduled admissions and the impact of variability in the number of admissions on the required bed capacity. First we study approximations for determining the impact of the daily variability in the number of admissions, for both stationary and time-dependent admissions with a weekly pattern. This results in intuitive approximations for the variability in bed demand and for blocking probabilities. Second, we use these results in an optimisation model that minimises the weighted deviations of the load from a pre-defined target load, which can differ from day to day. We incorporate emergency arrivals, routing of patients over different wards and multiple patient types, each type having a specific phase-type LOS distribution. Our primary focus is on bed demand, where the surgery scheduling may be included as a constraint.

Patient scheduling has received quite some attention in the literature, mostly focusing on the scheduling of surgeries. An example of work that studies surgery scheduling in combination with bed usage is Beliën and Demeulemeester [3], who try to level the bed usage by finding the best allocation of OT time blocks to surgical disciplines. They view the number of patients admitted on a day as a stochastic variable with a distribution depending on the specialty that used the OT. Van Oostrum et al. [21] find the optimal so-called master surgical schedule, in which they schedule all regularly performed surgeries on a specific day in the planning cycle, with a combination of OT time usage and the

maximum number of beds needed on every day as the objective function. They treat the length of stay as deterministic, with the length depending on the type of surgery performed. Vanberkel et al. [22] study the effect of a given surgical schedule on the usage of beds, taking emergency arrivals and different ward types into account as well. However, they do not use an optimisation algorithm and only try to improve step-by-step by trial and error. Their approach has been applied in practice with good results.

Gallivan and Utley [10] present a generic model for determining the distribution of bed occupation for a given cyclic admission schedule. They give an example of how these results could be used in an optimisation context. They restrict themselves to a single ward. Adan et al. [1] present a case study in which they apply an optimisation model. They consider both the OT usage and several other types of resources, such as different wards visited by patients consecutively. A weighted combination of the over- and underutilisation of all these resources is minimised, in both a deterministic and a stochastic version. The stochastic version cannot be solved to optimality due to its size in their case study setting, although they do believe that taking randomness into account is important.

The remainder of the paper is organised as follows. We start with quantifying the impact of variability on the required bed capacity. In Section 2 we use approximation methods for analysing models with non-Poisson arrivals and we analyse time-dependent arrivals, to allow for a weekly arrival pattern, in Section 3. In Section 4 we discuss admission scheduling that results in a stable bed demand by applying a Quadratic Programming model. We conclude with Section 5 where we discuss the contribution of this paper and describe the main practical insights that can be derived.

2 Impact of variability in scheduled admissions

The arrival process of emergency admissions is generally well approximated by a Poisson process. Although elective admissions are scheduled, our experience is that the variability in the number of elective admissions is at least as large as the variability in the number of unscheduled admissions, which is also supported by various studies, see e.g. [5, 20]. Given the variability in both types of admissions, the Erlang loss (or delay) model is often well applicable for giving insight in the implications of capacity decisions for clinical wards, see for example [5, 19].

In this section we quantify the impact of a more stable (elective) arrival stream and the corresponding

appropriate capacity. Equivalently, this may be used to determine a target load in Section 4. We build on approximations in the literature to analyse models with a general stationary arrival process that is not necessarily Poisson. The approximations described here are further adjusted in Section 3, where we study systems with non-stationary arrivals.

For the Erlang loss model, the capacity is fixed at s beds. Patients are assumed to arrive according to a Poisson process with an average of λ per day. An arriving patient is admitted in case a bed is available and refused otherwise. An admitted patient stays for a stochastic duration (the length of stay) at the ward with an average of β days. By Little’s formula, the above implies that the offered load is $\rho := \lambda\beta$, which represents the average number of occupied beds in case there would always be sufficient capacity.

This $M/G/s/s$ model has been well studied. The probability that an arriving patient is refused, also called blocking probability, is then given by

$$B(s, \rho) = \frac{\rho^s/s!}{\sum_{k=0}^s \rho^k/k!}.$$

Moreover, the offered load (number of patients present in case of sufficient capacity) has a Poisson distribution, which can be well approximated by a normal distribution for ρ not too small. In particular, for the Poisson distribution the mean and variance are equal, which directly yields that the variance of the offered load can then be approximated by ρ .

To obtain insight in the impact of scheduled admissions it is required to eliminate the assumption of Poisson arrivals, which is crucial for most queueing models. This elimination leads to a $G/G/s/s$ queue, which is discussed next.

2.1 Stationary approximations

We approximate the $G/G/s/s$ queue using its infinite-server counterpart $G/G/\infty$. We assume a stationary arrival process, where arrivals occur at rate λ . The coefficient of variation of the interarrival time is denoted by c_a . The service times (lengths of stay) are assumed to be independent and identically distributed with mean β . We also introduce the so-called Gini coefficient, which surprisingly appears in the approximations. This measure is related to the Lorenz curve, which is used in economics to represent the inequality in the distribution of wealth or income among the citizens of a country. Here we use it for the inequality in the length of stay S among patients (see also [5]). The Gini coefficient is defined as the area under the Lorenz curve. For piecewise differentiable probability

distributions, the Gini coefficient (G), proposed in [7], is given by

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

For example, for a deterministic distribution we have $G = 0$, and for an exponential distribution $G = \frac{1}{2}$. In [5], the Gini coefficients are given for the LOS at different wards.

We start with an approximation for the number of busy servers, or rather the variance and distribution of the number of busy servers, for a $G/G/\infty$ system in heavy traffic. If we use the heavy-traffic approximation established in [4], we have that the number of busy servers X_ρ approaches a normal distribution in the limit when the load $\rho = \lambda\beta$ of the system tends to infinity:

$$\frac{X_\rho - \rho}{\sqrt{\rho z}} \rightarrow N(0, 1), \text{ as } \rho \rightarrow \infty,$$

with

$$\begin{aligned} z &= 1 + (c_a^2 - 1) \frac{1}{\mathbb{E}S} \int_0^\infty \mathbb{P}(S > y)^2 dy \\ &= 1 + (c_a^2 - 1) (1 - G), \end{aligned} \tag{1}$$

where the second equality follows directly from the representation of the Gini coefficient. We note that only the first equality seems to be available in the literature, and the interesting and useful relation to the Gini coefficient has not yet been observed. The z is a measure of the peakedness of the arrival process and the service times, see [24] for a more elaborate discussion. Here, the variance of the number of busy servers is $z\rho$. From the peakedness we can see that the variance increases with the squared coefficient of variation of the interarrival times c_a^2 as is to be expected, but it can either increase or decrease in the Gini coefficient depending on the sign of $(c_a^2 - 1)$. This means that reducing the variability in LOS is only beneficial in cases where the arrival process is already quite stable, implying that hospital managers should focus first on stabilising the arrival process before stabilising the LOS distribution. Note that the point at which $(c_a^2 - 1)$ changes signs corresponds to a Poisson arrival process.

The variability in offered load is of prime importance for the required amount of capacity. Based on the square root staffing rule, see e.g. [13, 25], the required number of beds is typically the mean offered load (ρ) plus a constant times the standard deviation in offered load ($\sqrt{z\rho}$). The latter term corresponds to buffer capacity to deal with variability in bed demand. The value of the constant depends on the service level target, but is often chosen to be between 1 and 2.

Table 1 The fraction of refused admissions for (mixed) stationary arrivals

Arrivals	LOS	Stdev. offered load	Loss fraction (%)
Det	Det	0	0.00
	Exp	3.42	2.51
	$H_2 (p_1 = 0.1)$	3.93	3.63
Det + Poisson	Det	3.42	2.51
	Exp	4.19	4.24
	$H_2 (p_1 = 0.1)$	4.41	4.75
Poisson	Any dist	4.84	5.80

The most natural performance measure for the G/G/s/s queue is the blocking probability B_c . In [24], the Hayward approximation is proposed, which is given by

$$B_c = B_c(s, \rho, z) \approx B\left(\frac{s}{z}, \frac{\rho}{z}\right). \quad (2)$$

In other words, we use the regular Erlang loss formula, but first divide both the number of servers and the load of the system by the peakedness z . This requires an extension of the Erlang loss formula to non-integer values for the number of servers, see [15].

From this approximation, it follows that the fraction of blocked arrivals increases as the peakedness increases. So, the fraction of blocked arrivals increases with the squared coefficient of variation of the inter-arrival times but can either increase or decrease with the Gini coefficient depending on the variability in the arrival process.

2.2 Numerical examples

Our aim with the numerical experiments is to obtain insight in the impact of elective admissions on the bed occupancy in hospitals. Since the Hayward approximation is available in the literature, it is not our goal to carry out extensive numerical experiments. As a base example, we consider an average-sized ward with 28 beds, see [5]. We present experiments with three different distributions for the length of stay, all with mean 4. The LOS at clinical wards can typically be represented by exponential or hyper-exponential distributions, whereas the deterministic LOS is included to obtain insight in the impact of the LOS characteristics. We consider (mixed) deterministic and Poisson arrivals, representing scheduled and emergency admissions. The average number of arrivals per day is 41/7, giving an average offered load of about 23.43.

The standard deviation in offered load and the fraction of refused admissions (blocking probability) for the different scenarios may be found in Table 1. These

values have been calculated using approximations and have been verified by simulation. We see that the standard deviation in offered load and the loss percentage increase with the share of Poisson arrivals.

Health chains In some scenarios, patients visit a number of successive wards before leaving the clinic. Heavy-traffic approximations for such networks are complicated, see [11, 23] for some extensions to networks. For some cases, the variability of downstream wards can easily be identified assuming sufficient capacity. For Poisson arrivals (unscheduled admissions) the number of patients in each node of the network has a Poisson distribution [17]. Furthermore, in case the LOS of the preceding wards are deterministic (e.g. pre-surgery admissions), the variability in admissions of the downstream ward inherits the variability of the original arrival process.

To illustrate the impact of a very regular admission schedule on downstream wards, we consider a specific tandem of two infinite server queues. We only consider deterministic external arrivals that arrive at queue 1 with an average of 41/7 arrivals per day. Each patient moves from queue 1 to queue 2 after which he/she leaves the network. We focus on queue 2 and choose an average length of stay (ALOS) of 4 for this queue leading to an average offered load of 23.43, such that the results for queue 2 may be compared to those of Table 1. The standard deviations of the offered load for queue 2 can be found in Table 2, which were determined using simulations. Note again that Poisson arrivals would lead to a Poisson number of patients present, yielding a standard deviation of 4.84.

Clearly, the variability in offered load for queue 2 is considerably larger than for queue 1, except in the case with deterministic ALOS at queue 1. The results show that the standard deviation in offered load increases with both the ALOS of queue 1 and with the squared coefficient of variation of the service times of queue 2. It is interesting to note that a more variable service time at queue 1 compared to an exponential distribution,

Table 2 Standard deviation offered load to queue 2

Queue 1		Queue 2 LOS distribution		
ALOS	LOS distribution	Det	Exp	$H_2 (p_1 = 0.1)$
1	Exp	2.40	3.74	4.18
	$H_2 (p_1 = 0.1)$	2.33	3.75	4.16
3	Exp	3.62	4.11	4.38
	$H_2 (p_1 = 0.1)$	3.31	4.04	4.35
5	Exp	4.01	4.27	4.61
	$H_2 (p_1 = 0.1)$	3.76	4.16	4.47
1, 3, 5	Det	0	3.42	3.93

i.e. H_2 ($p_1 = 0.1$), may reduce the variability in offered load at queue 2, in particular as the ALOS gets larger. More practically, we see that the impact of a regular admission schedule rapidly fades out for wards further down the health chain unless the length of stay is more or less fixed for the upstream wards.

3 Impact of time-dependent admissions

In this section, we assume that the arrival process at a ward depends on the day of the week. This case is of particular interest in view of the schedule of elective patients at the OT. For example, it is well known that the number of arrivals is generally smaller during the weekend than on weekdays since hardly any elective procedures are scheduled during the weekend, see e.g. [5, 12]. The assignment of OT sessions to surgical disciplines typically also leads to differences in the number of arrivals.

3.1 Time-dependent approximations

We assume that there is a periodic (cyclic) arrival pattern. Let T be the length of a cycle and denote the average number of arrivals during $[a, b]$ by $\lambda(a, b)$, $a \leq b$. We are mainly interested in the weekly pattern, i.e., $T = 7$. Let $\lambda_d = \lambda(d - 1, d)$ denote the average number of arrival at day d , $d = 1, \dots, 7$, where we denote Monday by day 1. Also, let $\bar{\lambda}$ be the average number of arrivals per day. As in Section 2, we assume that the capacity is fixed at s operational beds.

Again, we use the infinite server queue (G/G/∞) as a basis for approximating the number of occupied hospital beds. In particular, the mean number of occupied beds is $\rho = \bar{\lambda}\beta$ and the variance (in heavy-traffic) is $z\bar{\lambda}$, where z is called the peakedness reflecting the variability in arrival and LOS processes. Similar to [14], we assume that the variability in arrivals consists of a random and predictable part. We decompose the peakedness into a random and predictable part as well, yielding

$$z = z_{\text{rand}} + z_{\text{pred}} = 1 + (c_a^2 - 1)(1 - G) + z_{\text{pred}}, \tag{3}$$

where the first part (z_{rand}) is the same as in Section 2. We note that the fraction of refused admissions can be determined again using the Hayward approximation, see Section 2. It easily follows that the loss fraction becomes larger for time-dependent arrivals (compared to a stationary arrival process) due to the increased peakedness.

We determine the second part (z_{pred}) of Eq. 3 based on a deterministic fluid approximation, see [17]. Specifically, let $z_{\text{pred}} = \text{Var}[m(t)]/\mathbb{E}[m(t)]$ with $m(t)$ the mean number of busy servers at time t in the G/G/∞ queue:

$$m(t) = \mathbb{E} \left[\int_{t-S}^t \lambda(s) ds \right] = \int_0^\infty \lambda(t-s) \mathbb{P}(S > s) ds, \tag{4}$$

where $\lambda(s)$ is the arrival rate at time s (see e.g. [17]). From Eq. 4 we see that the mean number of occupied beds depends on the full distribution of S , the LOS. That is, the ALOS or first two moments of the LOS distribution are not sufficient to determine the mean number of occupied beds at a particular point in time.

Of prime interest in the present setting is the case $T = 7$ with arrival rates λ_d , $d = 1, \dots, 7$ and an exponential (or hyper-exponential) LOS distribution. The case of an exponential LOS distribution can be used as a building block for more involved service time distributions and (feed forward) networks, see Appendix.

Exponential LOS For later use, we indicate $m^{\text{exp}}(t)$ for the mean number of busy servers in case of exponential service times. For convenience, consider the time instants $d = 1, 2, \dots, 7$ corresponding to the end of each day. Then, using Eq. 4, we directly obtain the recursive relation, for $d \in \mathbb{N}$,

$$\begin{aligned} m^{\text{exp}}(d) &= \int_{s=0}^1 \lambda_d e^{-\mu s} ds + \int_1^\infty \lambda(d-s) e^{-\mu s} ds \\ &= \frac{\lambda_d}{\mu} (1 - e^{-\mu}) + e^{-\mu} \int_0^\infty \lambda(d-1-u) e^{-\mu u} du \\ &= \frac{\lambda_d}{\mu} (1 - e^{-\mu}) + e^{-\mu} m^{\text{exp}}(d-1), \end{aligned} \tag{5}$$

where the final step follows from Eq. 4, see also [2].

Using the above relation n times, we have

$$m^{\text{exp}}(d) = \frac{1}{\mu} (1 - e^{-\mu}) \sum_{i=0}^{n-1} \lambda_{d-i} e^{-\mu i} + e^{-\mu n} m^{\text{exp}}(d-n).$$

Taking $n = T$ and using the periodicity of the arrival rate and, hence, $m^{\text{exp}}(d) = m^{\text{exp}}(d - T)$, yields

$$m^{\text{exp}}(d) = \frac{1}{\mu} \frac{1 - e^{-\mu}}{1 - e^{-T\mu}} \sum_{i=0}^{T-1} \lambda_{d-i} e^{-\mu i}. \tag{6}$$

LOS as sum of exponentials Here, we assume that the LOS can be expressed as a sum of exponential terms: $\mathbb{P}(S > t) = \sum_{j=1}^J p_j e^{-\mu_j t}$. This directly applies to

hyper- and hypoexponential LOS distributions. For the former, we have $0 < p_j < 1$ and $\sum_{j=1}^J p_j = 1$, whereas the tail distribution for the latter is given by Eq. 17. Note that these cases may be equivalently interpreted as parallel and tandem networks where the LOS in each node is exponentially distributed. The hypoexponential case may be primarily applied for modelling series of subsequent wards, whereas the hyper-exponential distribution often provides a better fit for the LOS distribution compared to the exponential.

For the mean number of occupied beds, we have

$$m(t) = \int_{v=0}^{\infty} \lambda(t-v) \left(\sum_{j=1}^J p_j e^{-\mu_j v} \right) dv$$

$$= \sum_{j=1}^J p_j m_j^{\text{exp}}(t).$$

Now, the predictable variation in the number of occupied beds is approximated by

$$z_{\text{pred}} = \frac{1}{T-1} \sum_{d=1}^T (m(d) - \bar{m})^2 / \bar{m}, \tag{7}$$

where $\bar{m} = \sum_{d=1}^T m(d) / T = \lambda\beta$ is the average occupancy; for instance, $m(d)$ is given by Eq. 6 for exponential LOS. We note that this may seem involved at first glance, but z_{pred} may easily be computed in e.g. a spreadsheet. Moreover, this derivation provides the basis for scheduling elective admissions as presented in Section 4.

Remark 1 A different approach in case of time-dependent arrivals is the stationary process approximation, see [18]. The main idea of that approach is to capture the additional variability in the arrival process in c_a^2 , i.e., the time-dependent process is approximated by a stationary process that is more variable. The disadvantage of this approach is that the impact of the service time (LOS) distribution on non-stationary arrivals cannot always be properly taken into account.

3.2 Numerical experiments

To verify the modified peakedness approximation (Eq. 3), we consider the following modification of the clinical ward introduced in Section 2.2. The average number of arrivals during weekdays and during the weekend is assumed to be 7 and 3, respectively. The ALOS is 4 days again, yielding an average load of roughly 23.43. For now, we assume the number of operational beds fixed at 28. In Table 3, we present approximation and simulation results for the standard deviation in offered load and the fraction of refused admissions for different LOS distributions and for both a deterministic and a Poisson arrival process. As can be observed from Table 3, the approximations are quite similar to the simulation results, indicating that the modified peakedness approximation (Eq. 3) works well.

Clearly, the weekly arrival pattern leads to increased variability in offered load and refused admissions compared to stationary arrivals (Table 1). This weekly pattern is most prominent for a deterministic LOS [2, 6]. As a consequence, the impact of the variability in LOS distribution on the offered load can go either way depending on the type of arrival process. This further strengthens the conclusion from Section 2 that the arrival process is of primary importance for a stable bed occupancy and should be considered first before focusing on the variability in LOS.

Health chains In Section 2 we illustrated that the benefits of a relatively stable arrival process to the first ward rapidly fades out for downstream wards due to the variability in LOS (except for deterministic LOS). Here, we give an example of the opposite effect in case of time-dependent arrivals. In particular, we consider two wards in tandem with sufficient capacity (infinite number of servers) in which the arrival process to the first ward is as described above. The LOS at both wards is exponentially distributed, where the LOS at ward 2 equals 4. The weekly pattern in average offered load

Table 3 The fraction of refused admissions for time-dependent arrivals

Arrivals	LOS	Stdev. offered load		Loss fraction	
		Approx	Simulation	Approx (%)	Simulation (%)
Det	Det	3.60	3.21	2.89	2.44
	Exp	3.91	3.80	3.58	3.36
	$H_2 (p_1 = 0.1)$	4.31	4.26	4.52	4.10
Det + Poisson	Det	5.00	4.66	6.18	6.09
	Exp	4.61	4.54	5.24	5.22
	$H_2 (p_1 = 0.1)$	4.76	4.63	5.61	5.52
Poisson	Det	6.03	5.76	8.76	9.6
	Exp	5.20	5.11	6.67	6.75
	$H_2 (p_1 = 0.1)$	5.15	5.05	6.57	6.71

can be found in Fig. 1 for an ALOS of 1, 3 and 5 days at ward 1. We also included the case in which the LOS at ward 1 equals 0, meaning that there effectively is a single ward. The different ALOS of ward 1 has the following implications for ward 2: (i) the peak in offered load shifts due to the LOS at ward 1, and (ii) the difference in average offered load across the week becomes smaller as the ALOS at ward 1 increases. Combined with the observations from Section 2, we note that, for non-deterministic LOS, the arrival process to downstream wards tend to look more like a stationary Poisson process.

4 Scheduling elective admissions

In most hospitals patient admission scheduling is done per medical discipline and independently of possible effects on the bed occupancy at clinical wards or intensive care units. As indicated, this often leads to high variability in bed occupancy and weekly patterns in the number of patients present at wards. The latter is also caused by the reduced number of (elective) admissions during the weekend (see Section 3). One way to deal with this weekly pattern is to adapt the staffing according to the offered load as in, e.g., [2] or [8]. A different approach is to schedule admissions such that undesired predictable fluctuations in the bed occupancy are avoided as much as possible. In this section, we propose a quantitative method for the latter option.

Specifically, the scheduling of elective admissions is done in two steps:

Step 1 Determine target load for each day, $m^*(d)$, $d = 1, \dots, T$.

Step 2 Determine an admission schedule such that the difference between the offered load and target load is minimised, using an optimisation model.

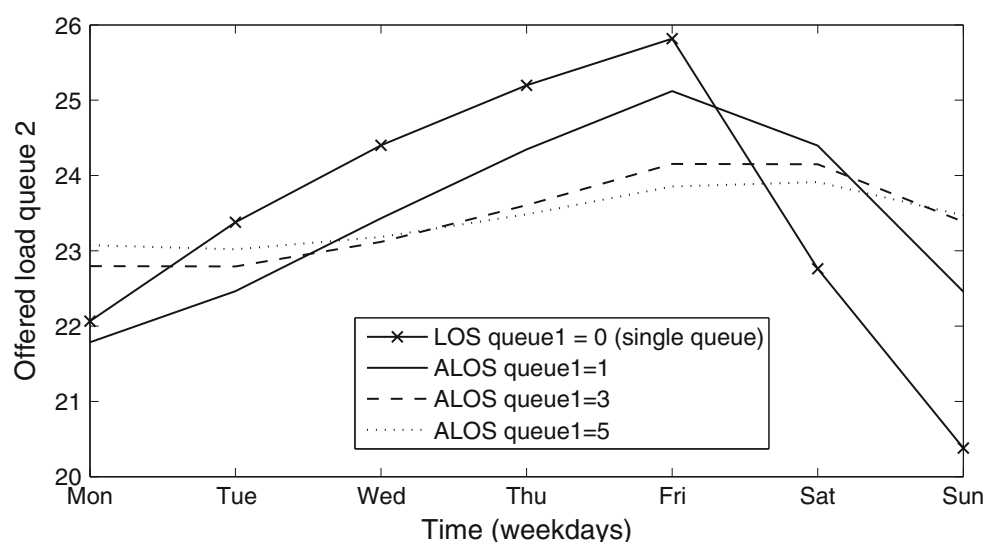
We note that Sections 2 and 3 play a key role in Steps 1 and 2, respectively. Here, we restrict ourselves to a single ward and K types of patients. This may, for instance, represent a ward for one medical discipline with various procedures leading to structural differences in the length of stay. This is just a base example and the model may be extended in various directions along the same lines. For implementation purposes, focusing on a single medical discipline may be a good starting point, as admissions are now generally scheduled per medical discipline and coordination between disciplines is not yet common. Moreover, in case the offered loads of all disciplines are well balanced, this immediately holds for the overall offered load.

We assume that the length of stay of patient type k is exponentially distributed with ALOS $1/\mu^k$. (Here we use exponential LOS, for hypo- or hyper-exponential LOS one ward is represented by more than one node.) Let the average number of admissions on day d of type k be λ_d^k and let the offered load of patients of type k on day d be $m^k(d)$. The target number of admissions for type k during T days is denoted by Λ^k .

Step 1: Target load

Determining the target load mainly concerns a managerial decision at a tactical level. It involves two parts: (i) The capacity in relation to variability in offered load, and (ii) the weekly pattern for available number of beds. Regarding (i), the models discussed in Section 2 can be applied to support decisions related to the

Fig. 1 The average offered load across the week for ward 2 in a tandem with different ALOS at ward 1



trade-off between occupancy levels and blocking probabilities. More specifically, the average load per day m^* may be determined using Eq. 2. For instance, for a given throughput the required capacity may be determined such that the blocking probability does not exceed some target. Alternatively, given a fixed capacity, a target occupancy level may be determined such that the blocking probability does not exceed some chosen value.

The number of available beds depends on the staffing, which is not necessarily the same for every day of the week. A typical example for (ii) is a different staffing during weekdays compared to the weekend, which generally means that during the weekend some beds are closed due to reduced bed demand (that is a consequence of the limited number of scheduled admissions). Denote the target load on day d by $m^*(d)$, $d = 1, \dots, T$. Clearly, it should hold that $m^* = \sum_d m^*(d)/T$. The target load during a cycle should also be equal to the offered load following from the admission target and corresponding ALOS, i.e., $m^*(d)$ should satisfy

$$m^* = \frac{1}{T} \sum_{d=1}^T m^*(d) = \frac{1}{T} \sum_{k=1}^K \Lambda^k \times \frac{1}{\mu^k}.$$

For identical targets on all days, we evidently have $m^*(d) \equiv m^*$. In case the number of open beds during the weekend is reduced by x (assuming that T is a multiple of 7), it follows after some straightforward calculations that

$$m^*(d) = m^* + \frac{2}{7}x \quad \text{for } d \text{ multiple of } 1, \dots, 5$$

$$m^*(d) = m^* - \frac{5}{7}x \quad \text{for } d \text{ multiple of } 6, 7$$

Step 2: Optimal admission schedule

In this step, we translate the admission scheduling into a mathematical model, using results from Section 3.

Specifically, we formulate the problem as a Quadratic Programming model with linear constraints, which is in the spirit of [1]. (We note that it can be formulated as a Linear Program as well using a different objective function.) The key element is that the time-dependent offered load as determined in Section 3 is linear in λ_d .

The objective here is to minimise the total squared deviation of the offered load from the target load. This is represented in Eq. 8 where the squared deviation between the target and offered load is summed over all days of the planning cycle.

The offered load for each patient type for the first (Eq. 11) and all consecutive days (Eq. 12) of the planning cycle are derived from the time-dependent analysis in Section 3, i.e., Eq 6 for exponential LOS. Note again that the full LOS distribution is required to determine the mean offered loads, and not just the average LOS. The total offered load on a particular day is the sum of the loads generated by the different patient types, as can be seen in Eq. 10.

The constraint 9 assures that the total number of scheduled admissions is equal to the target number of admissions for each patient type. The number of admissions on each day should be non-negative, as represented by Eq. 13. Moreover, it might be desirable or current practice that no scheduled admissions occur on some days, for instance, during the weekends or on days when no OT-time is available for a certain patient type. For such days, λ_d^k should be set to 0, as represented in Eq. 14.

Finally, we note that the choice of decision variables depends on whether patients of type k , $k = 1, \dots, K$, represent scheduled or unscheduled admissions. In case type k patients are scheduled, then λ_d^k , $d = 1, \dots, T$ are decision variables, whereas the λ_d^k should be determined from historical data in case of emergency admissions. (For the latter, $m^k(d)$ can also be determined directly from the data).

$$\text{Minimise } \sum_{d=1}^T [m(d) - m^*(d)]^2 \tag{8}$$

$$\text{subject to } \sum_{d=1}^T \lambda_d^k = \Lambda^k, \quad k = 1, \dots, K \tag{9}$$

$$m(d) = \sum_{k=1}^K m^k(d), \quad d = 1, \dots, T \tag{10}$$

$$m^k(1) = \frac{1}{\mu^k} \frac{1 - e^{-\mu^k}}{1 - e^{-T\mu^k}} \sum_{i=0}^{T-1} \lambda_{1-i}^k e^{-\mu^k i}, \quad k = 1, \dots, K \tag{11}$$

$$m^k(d) = \frac{\lambda_d^k}{\mu^k} (1 - e^{-\mu^k}) + e^{-\mu^k} m^k(d - 1), \quad d = 2, \dots, T, \quad k = 1, \dots, K \tag{12}$$

$$\lambda_d^k \geq 0, \quad d = 1, \dots, T, \quad k = 1, \dots, K \tag{13}$$

$$\text{(case-dependent) } \lambda_d^k = 0, \quad \text{for some } d \in \{1, \dots, T\}, \quad k \in \{1, \dots, K\} \tag{14}$$

As mentioned, the admissions scheduling can be modeled as a Linear Programming problem by modifying the objective function. In that case, the objective (Eq. 8) is to minimise $\sum_{d=1}^T |m(d) - m^*(d)|$, which can be made linear using standard LP arguments. Here we opt for a quadratic objective function because we assume that the consequences of a deviation from the target load will not be linear in the size of the deviation. It is considerably more difficult for the medical staff to handle larger deviations.

Extensions and modifications The QP model as introduced above is an elementary model that may be extended in different directions depending on the specific situation. Two important extensions are multiple (consecutive) wards and the impact of the Operating Theater for surgical patients. These extensions are discussed below.

A prime example where multiple wards are involved concerns medical disciplines for which a considerable fraction of the patients needs care at an ICU, after which they join the Normal Care Unit. The development of clinical pathways has also increased the interests in health chains. The time-dependent performance of health chains may again be found in Section 3 and Appendix, which is one of the key elements to extent the QP model, i.e., extend Eqs. 11 and 12. Moreover, the objective function should then be modified such that the sum of the deviations from the target load of each ward is minimised. Depending on its type, different wards may be assigned a different weight to represent its relative importance. For example, the weight for an ICU will typically be larger than the weight for other wards, as ICU capacity is more costly and the options for the transfer of patients in case of insufficient capacity is limited.

For surgical patients, the number of admissions is restricted due to the schedule of the OT. In general, each surgical discipline gets one or more rooms assigned to for some specific days, i.e., the OT sessions. For a given OT schedule, the maximum number of admissions of type k on some day d thus depends on the surgical time of type k patients and the available OT time for the medical discipline of type k . Such restrictions can be

straightforwardly included in the QP and thus easily allow for modifications in the admission planning without (strongly) affecting the OT schedule. Finally, we like to emphasise that the admission scheduling applies to both surgical and non-surgical patients.

4.1 Numerical experiments

The scenario for the numerical experiments is comparable to the scenarios considered in Sections 2.2 and 3.2. Specifically, emergency patients arrive with an average of 3 per day and have an ALOS of 4 days. For the elective patients, we assume that two groups can be distinguished: patients with short (ALOS of 2 days) and long (ALOS of 6 days) hospital stay. These groups can, for instance, be determined based on medical procedures. The target number of admissions for both groups is 10 patients per week. For simplicity, the length of stay is exponentially distributed for all groups.

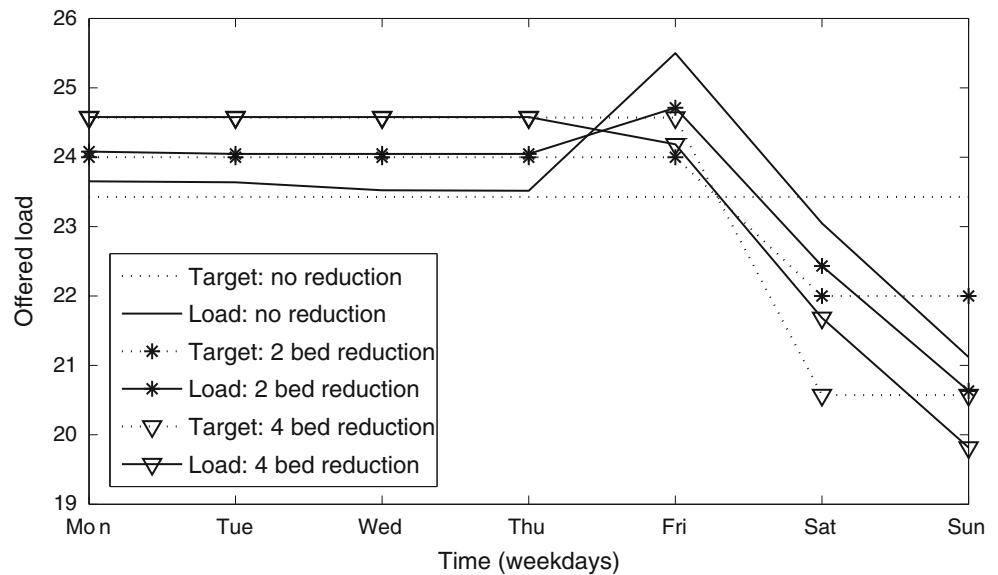
We consider three different target scenarios: no reduction of beds during the weekend and closing 2 and 4 beds during the weekend. For all scenarios, no elective admissions during the weekend are allowed. The required number of elective admissions that follow from solving the QP are given in Table 4. The resulting offered loads, along with the targets, are displayed in Fig. 2. We note that the presented number of admissions in Table 4 are fractional. To find the admission quota, these numbers could be rounded to the nearest integer. If it is infeasible to guarantee identical number of elective admissions for each patient group for a considerable time period, the “admission planner” could work with a small bandwidth. In practical situations the fractional numbers therefore provide a guideline in which direction the actual number of admissions should deviate from the prescribed admission quota.

Observe that in all scenarios, the target and offered load are not identical for all days of the week. Because there are no admissions during the weekend, there is limited control over the offered load during that period. For instance, in all cases the load decreases considerably from Saturday to Sunday. To compensate for the relatively small load on Sunday, the number

Table 4 Elective admission quota for different bed occupancy targets

Weekend reduction		Mon	Tue	Wed	Thu	Fri	Sat	Sun
No weekend reduction	Short	5.2	3.6	1.2	0	0	0	0
	Long	0	0	2.2	3.0	4.7	0	0
Reduction of 2 beds	Short	6.2	3.8	0	0	0	0	0
	Long	0	0.1	3.6	3.0	3.3	0	0
Reduction of 4 beds	Short	4.0	3.2	0.5	1.0	1.1	0	0
	Long	3.3	0.3	2.9	2.1	1.4	0	0

Fig. 2 The target load and offered load across the week for different target scenario's



of admissions is largest on Mondays for all three scenarios. In case the reduction in beds is limited (here 0 or 2), the patients with relatively long hospital stay should be admitted at the end of the week, often on Fridays, thereby filling the beds during the weekend. As a consequence, the patients with short hospital stay are mainly admitted at the beginning of the week, often on Monday.

We note that aiming for a constant bed occupancy target might be undesirable in this example. For the scenario of no weekend reduction, the offered load has a peak on Friday that is implicitly caused by the relatively large target during the weekend. In case 2 beds are closed during the weekend, there remains a smaller peak in offered load on Friday, whereas this peak is absent in the scenario where 4 beds are closed. Although, we presented a specific case, the admission principles apply to a broader health care setting.

5 Practical implications and discussion

The main goal of this paper is to provide quantitative methods to determine admissions quota for scheduled admissions and to analyse the impact of variability in scheduled admissions on the required bed capacity. For the impact of variability, we used approximation methods that build on heavy-traffic results in the literature and presented an interesting relation to the Gini coefficient. Moreover, we modified this peakedness approximation to allow for time-dependent arrivals, which is exploited in the step of admission schedul-

ing. In particular, the admission quota for scheduled patients are determined using a QP model minimising the difference between the expected and desired occupancy.

Our second aim is to derive generic practical insights that apply to almost all hospital situations. A first major observation is that more variation in admissions leads to a higher variability in bed demand and to more refused admissions for a hospital ward. Variation in the LOS can have negative consequences as well, but its influence depends on the variability of the arrivals. Only for stable arrival processes, reducing the variation in LOS leads to a less variable bed occupancy. Hence, stabilising the bed occupancy is best achieved by starting to smoothen the admissions. Along the same lines, the most time-stable performance is achieved when the arrivals to the hospital are as evenly distributed over the week as possible. A very uneven weekly pattern will increase variability in bed demand and the probability of refused admissions just as a variable arrival process will. If there is a clear weekly pattern, a LOS that is very stable can even be detrimental. Here, again, one should start by smoothing the admission pattern.

In practice, patients often visit more than one type of ward during their stay in the hospital. The variation in demand at the first ward influences that on subsequent wards. For relatively stable admissions, the variability in bed demand on the second ward is at least as large as that on the first one. Typically, for subsequent wards the bed occupancy starts to look more like the occupancy generally seen for emergency patients. In situations with a weekly admission pattern, a peak in demand on the first ward will be noticeable

for the second ward as well, but with a shift in time on which it occurs. The weekly pattern on the second ward becomes less noticeable as the LOS in the first ward becomes more variable.

In addition to smoothing the arrival process, it is also possible to schedule the arrivals in a better way. The first decision needed is the number of beds that will be staffed every day of the week, e.g., how many beds are closed during the weekend. In general, in absence of scheduled admissions during the weekend, it is advisable to close beds during that period. The case that no beds are closed during the weekend and scheduled admissions are absent, might lead to unused capacity. Scheduling patients with a longer expected LOS on Fridays can help to minimise this unused capacity, as such patients will stay throughout the weekend. The optimal schedule generated by our optimisation model typically generates schedules with such patterns (although the pattern is clearly affected by the number of closed beds during weekends). The drawback is that this often results in peak demand on Friday itself.

Another general rule is that more admissions should be scheduled on Mondays compared to the other weekdays, to fill the ward after the weekend. Because patients with a larger LOS are mainly scheduled at the end of the week, the patients scheduled for Mondays typically have a shorter LOS. Tuesday through Thursday are often comparable and roughly have about the same number of admissions scheduled.

We like to stress that the models presented are of a generic nature and can easily be implemented in e.g. an Excel spreadsheet to model the characteristics of a specific ward or hospital (or a specific time scale). Such models are of a deductive nature, based on a set of general principles and logical inference to derive new insights or improve decision making, see also Gallivan [9] for a further discussion on the role of models in health care. By definition, these models are based on assumptions regarding the structural characteristics of patient flows and admissions and, therefore, do not capture all decisions made at a hospital. A particular topic that is not captured by the models presented here is the possible dependence of the discharge process on the day of week or the occupancy. Although it is not clear whether it is desirable to incorporate such dependencies in the structural organisation of health care processes, this provides an interesting topic for further research.

Acknowledgements The authors would like to thank the anonymous referees for their valuable comments.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Appendix: Phase type LOS and feed forward networks

In this section, we consider a feed forward network of nodes with an exponential LOS. This may be equivalently interpreted as a single node with a (specific) phase type LOS distribution. For convenience, we assume that the LOS rates are different for each node.

First, we define some notation, in line with Section 3, and restate part of a more general result of [17]. Let J be the number of nodes and let $\lambda_j(t)$ be the external arrival rate to node j at time t . A patient goes from node i to node j with probability p_{ij} and leaves the network from node i with probability $1 - \sum_{j \neq i} p_{ij}$. Denote a generic LOS at node j by S_j and let μ_j be its LOS rate.

The main goal is to determine the mean number of occupied beds at time t for node j ($m_j(t)$). We first restate (part of) a more general result of Massey and Whitt, see [17, Theorem 1.2]:

Theorem 1 *In the $(M_t/GI/\infty)^J/M$ model, the number of occupied beds $Q_j(t)$ at time t , $1 \leq j \leq J$, are independent Poisson random variables with finite means*

$$m_j(t) = \mathbb{E}[Q_j(t)] = \mathbb{E} \left[\int_{t-S_j}^t \lambda_j^+(u) du \right], \tag{15}$$

where λ_j^+ is the aggregate-arrival-rate function to node j , defined as the minimal nonnegative solution to the system of input equations, for $1 \leq j \leq J$,

$$\lambda_j^+(t) = \lambda_j(t) + \sum_{i=1}^J \mathbb{E} [\lambda_i^+(t - S_i)] p_{ij}. \tag{16}$$

For optimisation purposes and for applications in health care, we are interested in more explicit results. Therefore, we make several assumptions, while maintaining a sufficiently generic framework for modelling in practical situations. We assume that all S_j are exponential, i.e., we restrict ourselves to phase-type LOS distributions. For convenience, we also assume here that all μ_j 's are different. Finally, we consider a feed forward network meaning that $p_{ij} = 0$ for $j \leq i$ and $1 \leq i \leq J$.

Below, we express $m_j(t)$ in terms of single nodes with exponential LOS, which are essentially used as building blocks. To do so, we decompose the patient flows into all possible routes through the network (that have

nonzero probability). A patient on route r then uses a subset of the nodes $\{1, \dots, J\}$. Specifically, patients on route $r = \{n_1, \dots, n_f\}$ arrive at the first node n_1 with rate $\lambda_{n_1}(t)p_r$, where $p_r = p_{n_1 n_2} \cdots p_{n_{f-1} n_f}$ represents the fraction of traffic coming from node n_1 going through n_f via route r .

Now, consider node j and truncate the network at node j , i.e., consider the network consisting of nodes $\{1, \dots, j\}$. Let r^j be a route in the truncated network that goes through node j and let R^j be the set of possible routes going through j . We add a subscript s if route r^j starts at node s (we denote r_s^j and use R_s^j again to denote the set of possible routes). Using Eqs. 15, 16 and the feed forward structure, we have

$$\begin{aligned}
 m_j(t) &= \mathbb{E} \left[\int_{t-S_j}^t \left(\lambda_j(u) + \sum_{i=1}^j \mathbb{E}[\lambda_i^+(u - S_i)] p_{ij} \right) du \right] \\
 &= \mathbb{E} \left[\int_{t-S_j}^t \lambda_j(u) du \right] + \mathbb{E} \left[\int_{t-S_j}^t \sum_{i=1}^{j-1} p_{ij} \mathbb{E}[\lambda_i^+(u - S_i)] du \right] \\
 &= m_j^{\text{exp}}(t) + \mathbb{E} \left[\int_{t-S_j}^t \sum_{r_s^j \in R_s^j} p_{r_s^j} \mathbb{E} \left[\lambda_s \left(u - \sum_{l \in r_s^j} S_l \right) \right] du \right] \\
 &= m_j^{\text{exp}}(t) + \sum_{s=1}^{j-1} \sum_{r_s^j \in R_s^j} p_{r_s^j} \mathbb{E} \left[\int_{t-S_j}^t \lambda_s \left(u - \sum_{l \in r_s^j} S_l \right) du \right],
 \end{aligned}$$

where the final step follows from interchanging integrals and sum. Note that the value of the expectation is similar to the mean load in node j for a tandem network (or a single node with a hypoexponential LOS). Using the tail distribution of a hypoexponential random variable (Eq. 17), we get, after some rewriting,

$$\begin{aligned}
 &\mathbb{E} \left[\int_{t-S_j}^t \lambda_s \left(u - \sum_{l \in r_s^j} S_l \right) du \right] \\
 &= \mathbb{E} \left[\int_{t-(\sum_{l \in r_s^j} S_l + S_j)}^t \lambda_s(u) du \right] - \mathbb{E} \left[\int_{t-\sum_{l \in r_s^j} S_l}^t \lambda_s(u) du \right] \\
 &= \prod_{l \in r_s^j - \{j\}} \frac{\mu_l}{\mu_l - \mu_j} m_{\{\lambda_s(\cdot), \mu_j\}}^{\text{exp}}(t) \\
 &\quad + \sum_{i \in r_s^j - \{j\}} \frac{\mu_i}{\mu_j - \mu_i} \prod_{l \in r_s^j - \{j\}} \frac{\mu_l}{\mu_l - \mu_i} m_{\{\lambda_s(\cdot), \mu_i\}}^{\text{exp}}(t),
 \end{aligned}$$

where $m_{\{\lambda_s(\cdot), \mu_j\}}^{\text{exp}}(t)$ is the mean load at time t for a single node with exponential LOS at rate μ_j and arrival rate function $\lambda_s(\cdot)$. Combining the above yields

$$\begin{aligned}
 m_j(t) &= m_j^{\text{exp}}(t) + \sum_{s=1}^{j-1} \sum_{r_s^j \in R_s^j} p_{r_s^j} \\
 &\quad \times \left(\prod_{l \in r_s^j - \{j\}} \frac{\mu_l}{\mu_l - \mu_j} m_{\{\lambda_s(\cdot), \mu_j\}}^{\text{exp}}(t) \right. \\
 &\quad \left. + \sum_{i \in r_s^j - \{j\}} \frac{\mu_i}{\mu_j - \mu_i} \prod_{l \in r_s^j - \{j\}} \frac{\mu_l}{\mu_l - \mu_i} m_{\{\lambda_s(\cdot), \mu_i\}}^{\text{exp}}(t) \right).
 \end{aligned}$$

Again, the mean offered load at some node may seem involved at first glance, but we can still express it in terms of single exponential nodes. For a feed forward network of size J we require the time-dependent analysis of at most $J!$ exponential single nodes (for $j = 1, \dots, J$, we need $m_{\{\lambda_s(\cdot), \mu_j\}}^{\text{exp}}(\cdot)$, with $s = 1, \dots, j$). However, the actual required number of single exponential nodes strongly depends on the routing probabilities in a specific practical situation and will often be much smaller than $J!$.

Example 1 An important special case is a tandem network of J nodes in series. Assuming here that all customers arrive at the first node, the sojourn time is then the convolution of J exponentials, which has a hypoexponential distribution, i.e.,

$$\mathbb{P}(S > t) = \sum_{j=1}^J \prod_{n \neq j} \frac{\mu_n}{\mu_n - \mu_j} e^{-\mu_j t}. \tag{17}$$

In this case, the mean number of occupied beds in node j reads

$$\begin{aligned}
 m_j(t) &= \prod_{l=1}^{j-1} \frac{\mu_l}{\mu_l - \mu_j} m_j^{\text{exp}}(t) \\
 &\quad + \sum_{i=1}^{j-1} \frac{\mu_i}{\mu_j - \mu_i} \prod_{l \neq i}^{j-1} \frac{\mu_l}{\mu_l - \mu_i} m_i^{\text{exp}}(t).
 \end{aligned}$$

References

1. Adan I, Bekkers J, Dellaert N, Vissers J, Yu X (2009) Patient mix optimisation and stochastic resource requirements: a case study in cardiothoracic surgery planning. *Health Care Manag Sci* 12:129–141
2. Bekker R, de Bruin A (2010) Time-dependent analysis for refused admissions in clinical wards. *Ann Oper Res* 178:45–65

3. Beliën J, Demeulemeester E (2007) Building cyclic master surgery schedules with leveled resulting bed occupancy. *Eur J Oper Res* 176:1185–1204
4. Borovkov A (1967) On limit laws for service processes in multi-channel systems. *Sib Math J* 8:746–763
5. de Bruin A, Bekker R, van Zanten L, Koole G (2010) Dimensioning clinical wards using the Erlang loss model. *Ann Oper Res* 178:23–43
6. Davis J, Massey W, Whitt W (1995) Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Manag Sci* 41:1107–1116
7. Dorfman R (1979) A formula for the Gini coefficient. *Rev Econ Stat* 61:146–149
8. Feldman Z, Mandelbaum A, Massey W, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Manag Sci* 54:324–338
9. Gallivan S (2008) Challenging the role of calibration, validation and sensitivity analysis in relation to models of health care processes. *Health Care Manage Sci* 11:208–213
10. Gallivan S, Utley M (2005) Modelling admissions booking of elective in-patients into a treatment centre. *IMA J Manag Math* 16:305–315
11. Glynn P, Whitt W (1991) A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv Appl Probab* 23:188–209
12. Green L, Nguyen V (2001) Strategies for cutting hospital beds: the impact on patient service. *Health Serv Res* 36:421–442
13. Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper Res* 29:567–588
14. Holtzman J, Jagerman D (1979) Estimating peakedness from arrival counts. In: *Proceedings of ITC-9*. Torremolinos, Spain
15. Jagers A, van Doorn E (1986) On the continued Erlang loss function. *Oper Res Lett* 5:43–46
16. Litvak E, Buerhaus P, Davidoff F, Long M (2005) Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. *Joint Comm J Quality Saf* 31:330–338
17. Massey W, Whitt W (1993) Networks of infinite-server queues with nonstationary Poisson input. *Queueing Syst* 13:183–250
18. Massey W, Whitt W (1996) Stationary-process approximations for the nonstationary Erlang loss model. *Oper Res* 44:976–983
19. McManus M, Long M, Copper A, Litvak E (2004) Queueing theory accurately models the need for critical care resources. *Anesthesiology* 100:1271–1276
20. McManus M, Long M, Copper A, Mandell J, Berwick D, Pagano M, Litvak E (2003) Variability in surgical caseload and access to intensive care services. *Anesthesiology* 98:1491–1496
21. van Oostrum J, van Houdenhoven M, Hurink J, Hans E, Wullink G, Kazemier G (2008) A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum* 30:355–374
22. Vanberkel P, Boucherie R, Hans E, Hurink J, van Lent W, van Harten W (2010) An exact approach for relating recovering surgical patient workload to the master surgical schedule. *J Oper Res Soc*. doi:10.1057/jors.2010.141
23. Whitt W (1982) On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Adv Appl Probab* 14:171–190
24. Whitt W (1984) Heavy-traffic approximations for service systems with blocking. *ATT Bell Lab Tech J* 63:689–708
25. Whitt W (1992) Understanding the efficiency of multi-server service systems. *Manag Sci* 38:708–723