Health Care Manage Sci (2007) 10:125–137
DOI 10.1007/s10729-007-9009-8

Modeling the emergency cardiac in-patient flow: an application of queuing theory

Arnoud M. de Bruin · A. C. van Rossum · M. C. Visser · G. M. Koole

Received: 20 July 2006 / Accepted: 16 October 2006 / Published online: 20 April 2007

Abstract This study investigates the bottlenecks in the emergency care chain of cardiac in-patient flow. The primary goal is to determine the optimal bed allocation over the care chain given a maximum number of refused admissions. Another objective is to provide deeper insight in the relation between natural variation in arrivals and length of stay and occupancy rates. The strong focus on raising occupancy rates of hospital management is unrealistic and counterproductive. Economies of scale cannot be neglected. An important result is that refused admissions at the First Cardiac Aid (FCA) are primarily caused by unavailability of beds downstream the care chain. Both variability in LOS and fluctuations in arrivals result in large workload variations. Techniques from operations research were successfully used to describe the complexity and dynamics of emergency in-patient flow.

Keywords Length of stay · Capacity management · Queuing theory · Occupancy rate · Emergency patient flow

A. M. de Bruin (✉)
Division IV (room PK 6X.185),
VU University Medical Centre,
De Boelelaan 1117, PO Box 7075,
Amsterdam, The Netherlands
e-mail: am.debruin@vumc.nl

A. C. van Rossum
Department of Cardiology, VU University Medical Centre,
Amsterdam, The Netherlands

M. C. Visser
Department of Neurology, VU University Medical Centre,
Amsterdam, The Netherlands

G. M. Koole
Faculty of Sciences, Department of Mathematics, VU University,
Amsterdam, The Netherlands

1 Introduction

Capacity decisions in hospitals are made in general without the help of quantitative model-based analyses [12]. Over the past years hospital managers have been stimulated to reduce the number of beds and increase the occupancy rates to improve operational efficiency. This strategy is questionable. Variability in length of stay (LOS) has a major impact on day-to-day hospital operation and capacity requirements. If this variability is disregarded during modeling an unrealistic and static representation of reality will emerge. A model, only based on average numbers, is not capable of describing the complexity and dynamics of the in-patient flow. This is also known as the flaw of averages.

Management does not consider the total care chain from admission to discharge, but mainly focuses on the performance of individual units. Not surprisingly, this has often resulted in diminished patient access without any significant reduction in costs. The suggested solutions are suboptimal.

In this study we investigate the emergency in-patient flow of cardiac patients in a university medical centre. This particular patient flow is characterized by time-varying arrivals at the First Cardiac Aid (FCA), the department where emergency cardiac patients enter the hospital. After initial treatment patients are transferred to the Coronary Care Unit (CCU) before they go to the normal care clinical ward (NC).

Many hospitals have trouble keeping the right resources, such as beds and personnel, available for arriving patients. Measurements show that the CCU in the considered hospital operates at occupancy rates greater than 95%. As a result, it frequently occurs that the CCU has insufficient capacity because the unit is full. Consequently, the number

of refused admissions at the FCA is significant and numerous patients are turned away to other referring hospitals.

This is unacceptable and puts a great pressure on the required quality of care. More and more hospitals have to account for their quality of care. An admission guarantee for all patients entering the emergency department is one of the main goals of the hospital. Besides this service requirement, one has to consider the medical emergency aspect. In case of a heart attack, the sooner someone gets to the emergency room, the better his or her chance of not only surviving, but also of minimizing heart damage following the attack. This is often referred to as the 'Golden Hour' [14]. This study applies a queuing model to analyze congestion in the emergency care chain. With this model the number of beds in the care chain is determined for several service levels.

In Section 2 the structural model is constructed followed by the data analysis in Section 3. Section 4 describes the impact of fluctuations in arrivals and variation in LOS on capacity requirements. In Section 5 the phenomenon of blocking and the mathematical model are introduced. Section 6 gives the results and the paper ends with the conclusion and discussion in Section 7.

2 Structural model

The first phase of the study is the construction of a structural model (or flowchart) of the patient flow. Such a model describes the different patient routings in a qualitative manner and defines the relations between different hospital units. After expert meetings with cardiologists we decided to identify two different patient flows. The primary patient flow enters the system at the FCA and leaves the hospital after a stay at the CCU and NC. The different departments are defined as follows:

- First Cardiac Aid: A hospital unit intended to provide rapid diagnosis and initiation of treatment for subjects with acute symptoms probably due to cardiac disease (for example chest pain, syncope, palpitations, dyspnea)
- Coronary Care Unit: A hospital unit that is specially equipped to provide intensive care of patients with severe acute or chronic heart disease (for example acute coronary syndromes, arrhythmia, heart failure)
- Normal Care: A hospital unit equipped to provide non-intensive care to a particular group of patients, in this case patients with cardiac disease.

A secondary patient flow, originating from surrounding hospitals, enters the CCU and returns to other hospitals after treatment, thus bypassing the NC. These patients are hospitalized to have immediate percutaneous (or balloon) angioplasty (PTCA) [3]. This kind of treatment is referred to as top-clinical care. Only certified hospitals are allowed to perform this type of medical procedure.

The structural model with the two different patient flows is shown in Fig. 1.

Health care processes are characterized by a great uncertainty. A large variety of possible patient routings can be distinguished. If we investigate the different flows throughout the hospital in great detail the flowchart becomes like the path of a pinball. Therefore, Fig. 1 is not striving for completeness. Nevertheless, it is possible to

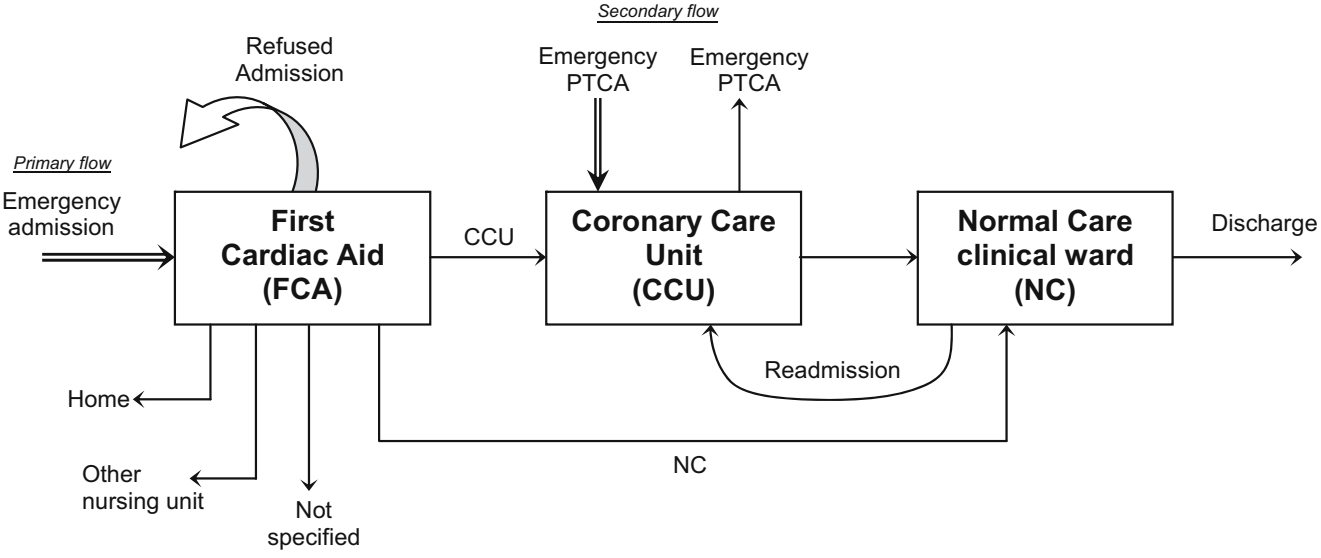

Fig. 1 Flowchart of the emergency cardiac in-patient flow

reduce complexity without losing integrity by focusing on the most critical patient flows. In this study both the primary and secondary patient flow are taken into account.

3 Data analysis

Computerized hospital records of all admissions at the FCA in 2003 have been used to analyze the arrival pattern of new patients. The patient flow is also quantified with the use of this data, in which the different routings and LOS distributions at the CCU and NC have been determined.

3.1 Arrivals

Historical data shows that the total number of annual arrivals (primary patient flow) fluctuates around 3,000. In 2003, the total number of arrivals at the FCA was 2,838. The average number of patients arriving per day is therefore 7.78. The unscheduled arrivals at the FCA are modeled as a Poisson process with intensity $\lambda=7.78$, see Fig. 2. The Poisson arrival assumption has been shown to be fitting in studies of unscheduled arrivals [19].

The FCA is characterized by time-varying arrivals during the day. Therefore, the arrival pattern over a 24 h period has been determined, see Fig. 3. Respectively, 14, 55 and 31% of all patients arrive in the intervals 00.00–08.00, 08.00–16.00 and 16.00–24.00.

As the LOS at the FCA (±6 h) is of the same order as the interval length in the arrival pattern (8 h) transient effects occur. The secondary patient flow is also unscheduled and modeled as a Poisson process with parameter $\lambda=1.37$.

3.2 Routings

In Section 2 the different patient routings are visualized in a flowchart. These routings are quantified in Table 1. Notice the alarming amount of refused admissions. In general, a high percentage of refused admissions is not uncommon for emergency departments [10]. Therefore, the organization of the emergency care is a hot topic for hospital professionals, managers and policy makers.

The refused admissions are calculated as a percentage of the total number of presentations at the FCA and are only observed in the primary patient flow.

The term ‘refused admission’ is to some extent misleading. All arriving patients are admitted to the FCA for initiation of treatment. After this first aid some patients have to be transferred to surrounding hospitals due to unavailability of beds downstream the care chain. This is what we entitle a refused admission.

3.3 Length of stay distributions

The number of days in hospital for a patient is described by the term length of stay (LOS). LOS is defined as the time of discharge minus time of admission. Following, the average length of stay is abbreviated as ALOS. In relation to industrial service times LOS distributions are characterized by a relatively high variability and a heavy tail. In probability theory the coefficient of variation (C_V) is a measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation (σ) to the mean (μ):

$$C_V = \frac{\sigma}{\mu}$$

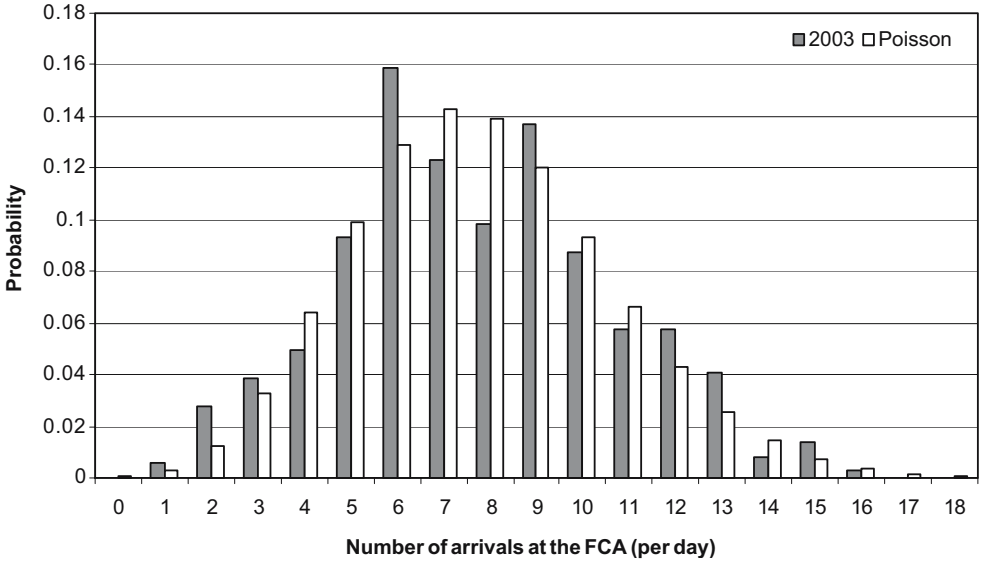

Fig. 2 Distribution of number of arrivals per day at the FCA, primary patient flow

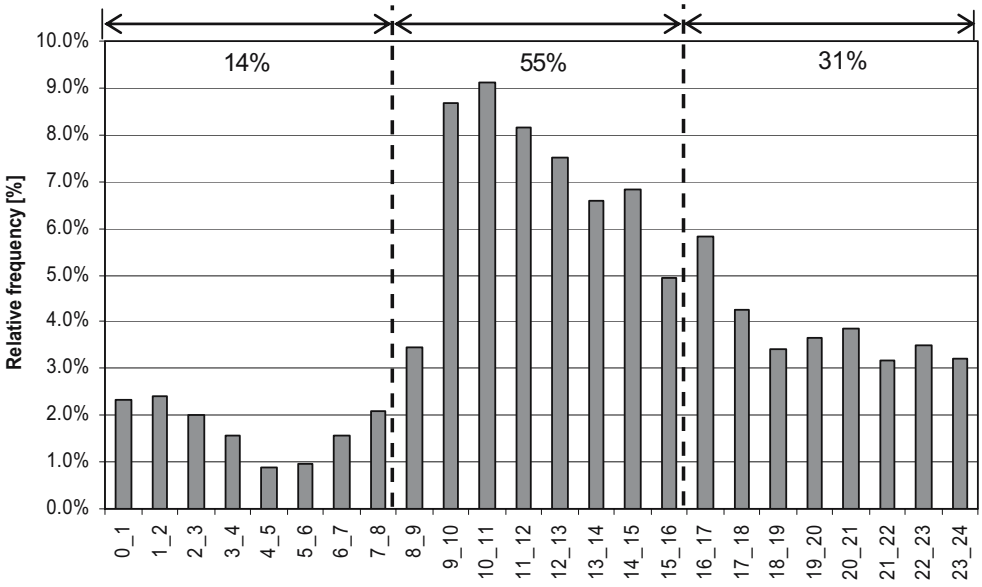

Fig. 3 24-h arrival pattern at the FCA

This coefficient is an important measure when describing health care processes. Values of the C_V are usually ≥1 [8]. For the exponential distribution the C_V equals 1. Another important factor is that the real (or measured) LOS is generally not equal to the LOS indicated by hospital professionals. Both medical and logistic reasons determine the LOS. The difference between the two, the additional LOS, is often caused by congestion or chain effects [11]. When a patient is ready to be transferred to another unit where no beds are available the patient remains at his current station and waits for a free bed. This type of congestion also occurs between hospitals and other health care institutions such as nursing homes. In other words, a certain part of the total LOS consists of additional time [2]. This fraction is often substantial; measurements indicate 20–30% of total LOS.

In the next three sections LOS-values are presented that were subtracted from the hospital information system, thus corresponding with the real LOS, including the additional time.

Table 1 Percentage routings at the FCA

Referral from FCA to:	2003	Percent
Home	1,899	66.9
Refused admission	383	13.5
Coronary Care Unit (CCU)	314	11.1
Normal care clinical ward (NC)	128	4.5
Other nursing unit	104	3.7
Not specified	10	0.4
Total	2,838	100

3.3.1 First Cardiac Aid (FCA)

For 2,401 (85%) of all arrivals the LOS was registered. The ALOS at the FCA is 6.4 h and patients never stay longer than 24 h. The median is 5 h and the C_V equals 0.7.

3.3.2 Coronary Care Unit (CCU)

At the CCU two types of patients can be identified. The first originates from the primary patient flow which enters the hospital at the FCA. The secondary patient flow consists of the emergency PTCA's (see Fig. 1). The ALOS of the primary patient flow at the CCU is 67 h, the median is 48 h and the coefficient of variation equals 0.99.

For the secondary patient flow, the ALOS is 18 h (median is 5), which is relatively short compared to the primary patient flow at the CCU. Nevertheless, the variability is remarkable (C_V=2.6). About 80% of all patients (group 1, N=394) leaves within the first 12 h of their hospital stay. The remaining 20% (group 2, N=100) has a prolonged hospital stay and occupy a relatively great part of the available resources. We define this demand as Total Resource Consumption (TRC) which is defined as the sum of all individual LOS values for both groups. The TRC of group 1 and group 2 is, respectively, 19 and 81%. Thus, we see that approximately 80% of the available resources are occupied by only 20% of the patients. This is known as Pareto's Principle or the 80/20 rule and is recognized in many quantitative studies. Due to the smaller volume of group two hospital professionals and management easily concentrate on group one. Thinking in terms of TRC group two is far more interesting as the potential gain in terms of

resource consumption, patient flow and consequently throughput is higher. Thus, the value of Pareto's Principle is that it reminds you to focus on the 20% that matters.

3.3.3 Normal Care clinical ward (NC)

The ALOS at the ward is 164 h, about 7 days. The median equals 113 h and the C_V is 1.07. As in the previous section two groups are identified. Group one, those patients with a LOS smaller than 240 h (10 days), contains 81% off the total number of patients, and consumes 47% of the available resources. Group two contains patients who stay longer than 10 days. This is the remaining 19% and they consume 53% of the capacity. The tail of the distribution is not as 'heavy' as at the CCU but the disproportional demand of both groups on the available resources remains remarkable. The LOS distribution at the NC is presented in Fig. 4.

Table 2 summarizes the LOS characteristics of the emergency care chain. As stated before these numbers can not be interpreted as constants of nature. The consequence of chain effects, such as congestion, on LOS-values must be considered seriously. The general conclusion is that health care processes are characterized by large variability in LOS. Coefficients of variation are often equal or greater than 1.

Many hospitals put energy in reducing the average length of stay. When a LOS-reduction has been realized at a particular nursing unit hospital management has two options:

1. Increasing the number of admissions on the same amount of beds.
2. Keeping the production targets constant while reducing the number of beds of the unit.

In both cases turnover per bed will rise. The choice depends largely on the type of contract hospitals make with health insurance companies.

Table 2 Summary of LOS-values in the emergency cardiac care chain

	Mean (μ) LOS (h)	Median LOS (h)	σ	C_V [σ/μ]
First cardiac aid	6.4	5	4.6	0.7
Coronary care unit, primary patient flow	67	48	66	0.99
Coronary care unit, secondary patient flow	18	5	47	2.6
Coronary care unit, mixed	44	22	62	1.4
Normal care	164	113	175	1.07

4 Impact of variation on capacity requirements

In this section the impact of fluctuations in arrivals and variation in LOS on capacity requirements is demonstrated. Queuing theory is used to quantify this impact. To demonstrate the effects we assume Poisson arrivals, exponential service times and an infinite number of beds. The outcome of this analysis is the number of beds required to accommodate all arrivals. This is relevant because one of the main goals of the hospital is providing an admission guarantee for all arriving patients. Nevertheless, in real life the number of beds is fixed and as a consequence patients

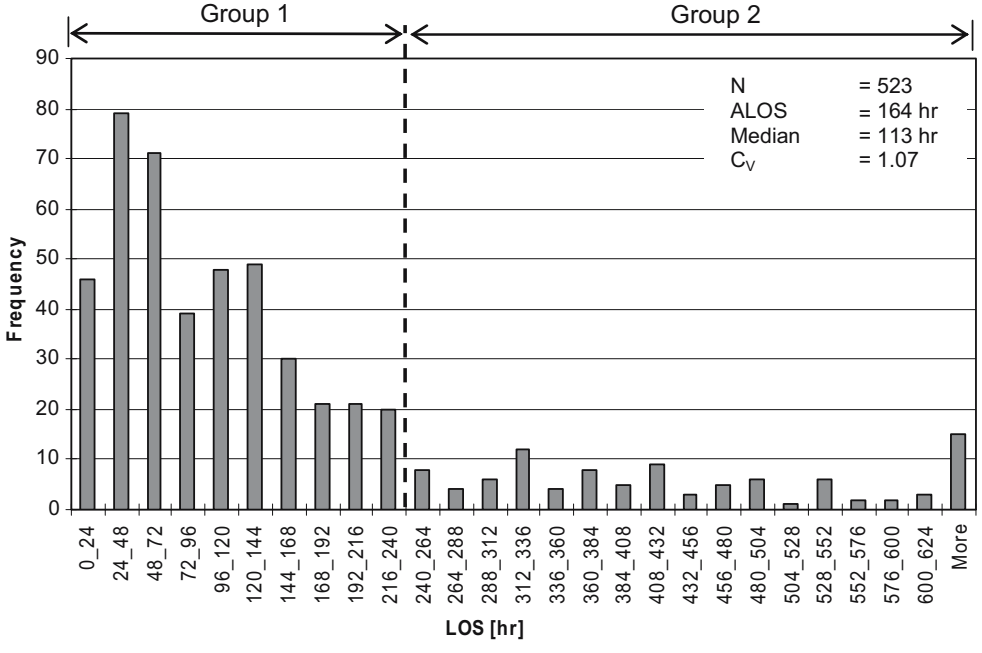

Fig. 4 LOS distribution at the NC

are turned away. This situation, with limited capacity and blocking, is described in Section 5.

As mentioned in Section 3 Poisson arrivals can be used to describe unscheduled hospital admissions [19] and for simplicity we assumed exponential service times. The queuing system under investigation is referred to as M\M\∞ in Kendall's notation [18]. The following parameters describe the queuing system:

λ average arrival rate
μ average length of stay
$B(t)$ number of patients in the system at time t or number of beds occupied at time t

A very important and powerful formula in the operations research, defined by Little [13] can be applied to almost every queuing system (Eq. 1). It shows the relation between the expected number of patients in the system, $EB(t)$, and the average length of stay (μ),

$$\mathrm{EB(t)} = \lambda\mu \tag{1}$$

Due to variations in number of arrivals and LOS the average value is exceeded on a regular basis. For example, at an intensive care unit (ICU) five patients arrive per day on average. The average LOS is 6 days. The parameters of this queuing system are: λ=5 and μ=6. Using Little's formula, the expected number of patients at the ICU is 30. If management decides to size the unit on this average based calculation, operational problems will occur on a regular basis. For the M\M\∞ model one can easily calculate the probability that i beds are occupied (Eq. 2). It is just a function of $EB(t)$, or $\lambda\mu$, the expected number of patients in the system,

$$\mathrm{P}_i = \mathrm{e}^{-\lambda\mu}\frac{(\lambda\mu)^i}{i!} \tag{2}$$

The probability that more than 30 beds are occupied is easily calculated,

$$\mathrm{P}(i > 30) = \sum_{i=31}^{\infty}\mathrm{P}_i = \sum_{i=31}^{\infty}\mathrm{e}^{-30}\frac{(30)^i}{i!} = 0.45$$

In other words, due to variability in arrivals and LOS, in 45% of the time more than 30 beds are required. Thus, the average based calculation is not feasible. This is entitled the flaw of averages. Gallivan et al. [5] demonstrates that a high degree of reserve capacity (up to 30%) is required to avoid high rates of operation cancellation due to unavailable beds downstream the care chain. This is a first illustration of the huge impact of variation on capacity requirements. The next sections describe this in more detail.

4.1 Multiple state analysis

In this section the occupation of the FCA is investigated. As stated in Section 3 the arrivals at the FCA are characterized by a strong fluctuation over a 24-h period. The probability distribution of the number of beds occupied at the FCA is calculated with the use of a M\M\∞-model. First, the average arrival rate (λ=7.78) is assumed to be constant over the day. The ALOS (μ) equals $6.4/2.4 = 0.27$ days. Thus, the expected number of beds occupied is 2.1.

Then, the 24-h period is divided in three intervals of each 8 h. This is a practical choice driven by the observed arrival pattern and by the working hour's schedule of personnel. Table 3 derives the values of the arrival rate for these different intervals (notated as λ*). The arrival rate during office hours (08.00–16.00) increases significantly compared to the situation in which λ is kept constant. This is an important conclusion.

Figure 5 presents the probability distributions for the number of beds occupied for λ=7.78 (mean) and λ*=12.8 (max). The expected number of beds occupied increases from 2.1 to 3.4 (+62%). Flexible staffing levels are a possible answer to this strong variation in workload [6].

The approach via the definition of multiple states is just a first rough method to gain insight in the effects of variation in arrival rate. Simulation is often useful to capture the impact of time variant arrival rates at emergency departments [1].

4.2 Steady state analysis

For both the CCU and the NC a steady state analysis has been performed. This means that fluctuations in arrival rate are neglected. The M\M\∞-model is used again to calculate the state probabilities.

Table 3 Definition of arrival rate for three 8 h intervals

Interval	Arrival Rate Constant over the Day			Three 8 h Intervals		
	# of Arrivals	%	λ	# of Arrivals	%	λ^*
00.00–08.00	946	33.3	7.8	397	14	3.3
08.00–16.00	946	33.3	7.8 (mean)	1,561	55	12.8 (max)
16.00–24.00	946	33.3	7.8	880	31	7.2
Total	2,838	100	7.8	2,838	100	7.8

*Equivalent value of λ for different intervals

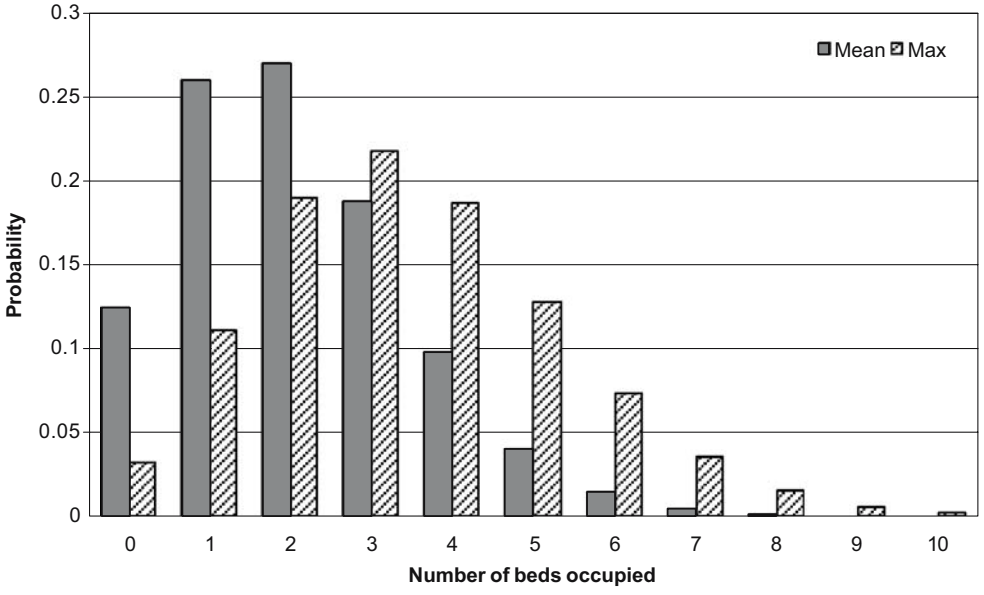

Fig. 5 The effect of fluctuation in arrivals on the FCA

The expected number of beds occupied is six for the CCU and 16 for the NC. The probability distribution of the number of beds occupied can be described with an exponential distribution. Therefore, using 6 and 16 beds, respectively, will result in day-to-day operational difficulties.

4.3 The situation in the hospital under investigation

In this section we compare the actual situation in the hospital with the quantitative analysis performed so far. The number of beds on the FCA and CCU is, respectively, 5 and 6. The NC is a mixed ward where both cardiothoracic and cardiac (scheduled and emergent) patients stay. The number of beds was 28 but as a consequence of the mixed population it is not possible to compare required and available resources for only the emergency cardiac in-patient flow.

The occupancy rate at the FCA is unknown but the analysis in Section 4.1 illustrates that bed capacity seems sufficient. Hospital professionals confirm this conclusion. In the average (λ=7.78) and maximum workload case (λ*= 12.8) the probabilities that the number of beds required exceeds five are, respectively, 2 and 13% (Fig. 5).

As stated in Section 4.2 we expect a CCU with six beds to have frequent operational difficulties. The occupancy rate distracted from management information is approximately 97%. This is consistent with the model and gives a good example of the flaw of averages in practice.

5 Modeling the emergency cardiac care chain

This chapter describes the model which has been developed for the primary goal of this study. In the previous section we assumed an infinite number of beds. In this section capacity is limited. First, in Section 5.1 the phenomena of blocking and economies of scale are introduced. Both blocking (e.g. refused admissions) and economies of scale are important features of health care processes and are directly related to the quality of care. Ridge et al. [15] also describe the non-linear relationship between number of beds, mean occupancy level and the number of patients that have to be transferred through lack of bed space. In Section 5.2 the queuing system which is used for this particular case is described shortly.

5.1 The phenomenon of blocking and impact of economies of scale on occupancy rates

An important model from queuing theory is the Erlang Loss model [4] or M\M\c\c in Kendall's notation. In this model customers (for example patients) arrive according to a Poisson process with intensity λ. This is the real demand, thus including the refused admissions. The LOS of arriving patients is independent and exponentially distributed with expectation μ. The number of beds is equal to c. There is no waiting area, which means that an arriving patient who finds all beds occupied is blocked. In real-life the consequence of blocking could well be a refused admission.

This is a more realistic representation of emergency in-patient flow. The fraction of patients which is blocked and sent away to other hospitals in the long run (P_c) can be calculated with the Erlang Loss formula,

$$P_c = \frac{(\lambda\mu)^c / c!}{\sum_{k=0}^{c} (\lambda\mu)^k \Big/ k!} \tag{3}$$

The occupancy rate (ρ) is related to the real demand (λ) and LOS (μ) and can be defined as follows,

$$\rho = \frac{\text{Average number of beds occupied}}{\text{number of beds available}} = \frac{(1 - P_c)\lambda \cdot \mu}{c} \quad (4)$$

The term $(1-P_c)\lambda$ can be entitled as the effective demand as the refused admissions are subtracted from the real demand. Furthermore, the product $\lambda\mu$ is known as the workload of the system.

Many hospitals use the same target occupancy rate for all hospital units, no matter the size of the unit. In general the unit size varies between 6 and approximately 60 beds. The target occupancy rate is typically set at 85% and has developed into a golden standard [7]. The feasibility of this target is no matter of discussion in the considered hospital.

In order to demonstrate the relation between the size of a hospital unit, the feasibility of the 85% target and the fraction of refused admissions two calculations have been made. Both calculations have been performed via iteration of Eqs. 3 and 4.

1. The percentage of refused admissions (P_c) given an occupancy rate (ρ) of 85% ($2 \leq c \leq 60$)
2. The target occupancy rate (ρ) for P_c=0.05 (5% refused admissions) ($2 \leq c \leq 60$)

Table 4 and Fig. 6 present the results both numerically and graphically.

The conclusion is clear and important. Larger hospital units can operate at higher occupancy rates than smaller ones while attaining the same percentage of refused admissions. Therefore, one target occupancy rate for all hospital units is not realistic. The 85% target is only attainable for units with more than 50 beds, assuming P_c= 0.05 is acceptable. If we hold the 85% target for a small unit such as the CCU (6 beds) nearly half of all arriving patients is blocked.

Currently, the discussion about refused admissions does not focus on the direct relation between the size of a hospital unit and the feasibility of target occupancy rates.

Table 4 Relation between number of beds, fraction refused admissions (P_c) and occupancy rates (ρ)

Number of Beds (c)	Refused Admissions (P_c) for ρ=0.85 (%)	Target Occupancy Rate for P_c=0.05 (%)
2	73.6	18.1
4	57.3	36.2
6	46.4	46.9
8	38.6	53.9
10	32.7	59.0
12	28.2	62.9
14	24.6	66.0
16	21.7	68.5
18	19.3	70.6
20	17.3	72.4
22	15.6	74.0
24	14.1	75.3
26	12.9	76.5
28	11.8	77.6
30	10.8	78.5
32	10.0	79.4
34	9.2	80.2
36	8.5	80.9
38	7.9	81.6
40	7.4	82.2
42	6.9	82.7
44	6.4	83.2
46	6.0	83.7
48	5.6	84.2
50	5.3	84.6
52	5.0	85.0
54	4.7	85.4
56	4.4	85.7
58	4.2	86.1
60	4.0	86.4

5.2 Two-dimensional Markov process

Simulation models have been frequently used to describe the emergency in-patient flow [1, 9, 16, 17]. Although simulation is a powerful tool for investigating complex systems, we belief the choice is often made arbitrarily and too easily. The complexity of the care chain in this study does not necessarily require simulation. Therefore, a two-dimensional Markov process with blocking is applied to analyze the congestion in the acute cardiac care chain. An analytical approach has several advantages over simulation:

- Costs are less
- Easier to implement
- Generates exact solutions

Two important reasons to choose for simulation models are:

- The graphical interface
- Simulation models are often more understood by doctors and managers and, therefore, more convincing

Furthermore, simulation can analyze the impact of time-varying arrival rates while a math model cannot. Also, simulation can be valuable when LOS data cannot be fit by any distribution. Nevertheless, we preferred an analytical approach for this particular case study.

The primary goal was to determine the optimal bed allocation over the emergency care chain. In the current

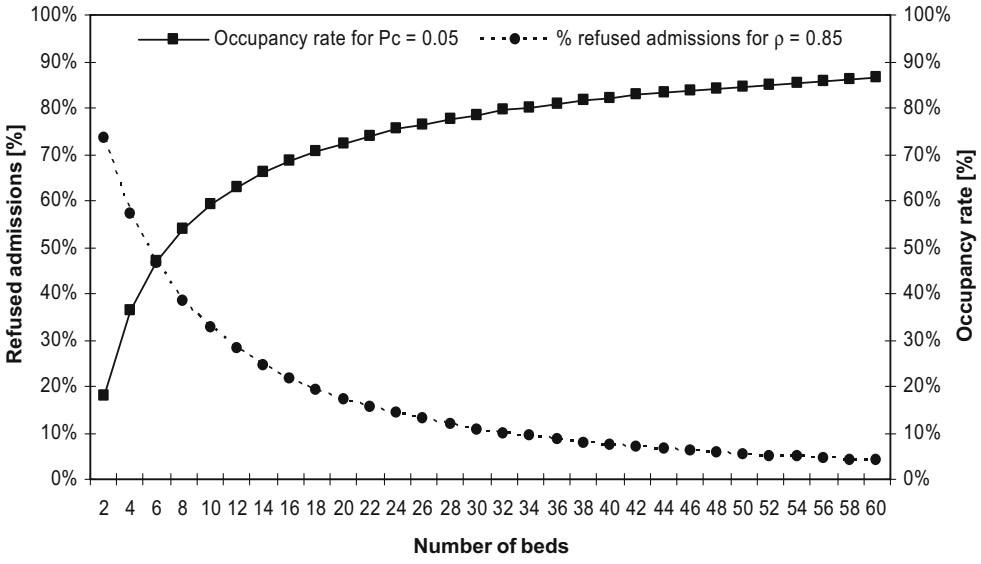

Fig. 6 Relation between number of beds, fraction refused admissions (P_c) and occupancy rates (ρ)

situation capacity at the FCA seems sufficient, and therefore, this station is left out of the analysis. This choice reduces the complexity of the model. The following parameters are introduced,

$$\left.\begin{array}{l} N_1 = number\ of\ beds\ at\ the\ CCU \\ N_2 = number\ of\ beds\ at\ the\ NC \end{array}\right\} Input\ variables$$
$$x = number\ of\ CCU - patients$$
$$y = number\ of\ NC - patients$$

with the following constraints,

$x + y \leq N_1 + N_2$ the total number of patients in the care chain is less than or equal to the total amount beds in the care chain

$x \leq N_1$ the number of CCU-patients is less than or equal to the number of beds at the CCU

A graphical representation of the problem is shown in Fig. 7. The connection line between x and N_1 means that CCU patients can only stay at the CCU. NC-patients (y) can stay in the CCU as well as in the NC which is visualized in Fig. 7 by the two lines starting at 'y'. Furthermore, there are three possible transitions:

1 Patient is transferred from the FCA to the CCU
2 Patient is transferred from the CCU to the NC
3 Patient is discharged from the NC

To reduce complexity the re-admissions (flowing from NC to CCU, see Fig. 1) are left out of the analysis.

A computer program has been written that computes numerically the stationary distribution of this two-dimensional Markov process. As output it gives the number of refused admissions for given arrival rate, number of beds at the CCU and NC and LOS at these stations.

6 Results of case study

In this section the numerical results of this study are presented. The primary goal is to determine the optimal bed allocation over the emergency cardiac care chain, given a required service level. Defining target service levels is

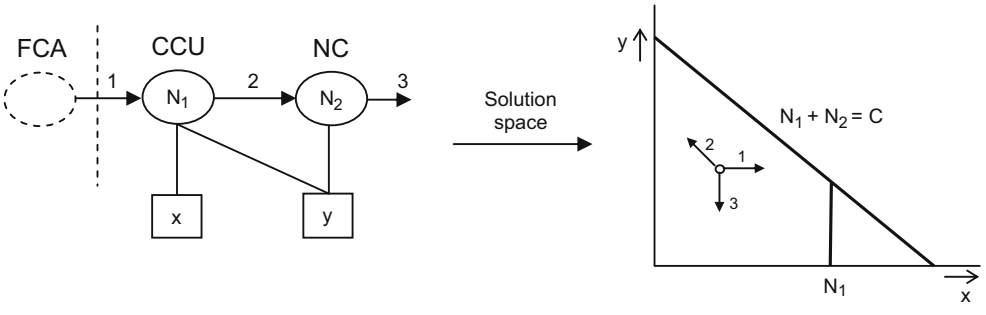

Fig. 7 Graphical representation of the problem

relatively new for healthcare institutions and originates from the service sector (e.g. call centers). The definition of what service level exactly means and how it is measured is critical. Service level in this study is closely related to the percentage refused admissions. The decision to refuse a patient will be influenced by patient characteristics, to what extent is unclear. The choice of how high the target service level must be is made arbitrarily.

The service level requirement is set at a maximum of 2% refused admissions at the FCA. This is a major improvement compared to the current fraction of refused admissions (13%). In our model, as described in Section 5.2, the blocking percentage at the CCU is calculated. Therefore, the service level requirement is rewritten into a maximum of 5% blocking at the CCU. In terms of number of refused patients that is equivalent.

A baseline measurement shows that the fraction of refused admission at the CCU equals $383/(383+314+500) = 32\%$. We relate the number of refused admissions to all arrivals at the CCU, thus including the secondary patient flow (500 pts). The outcome of this calculation of refused admissions is even more excessive than the calculation in Table 1. Approximately one out of three arrivals at the CCU is sent away.

We varied the number of CCU beds from 5 to 15 and the number of NC beds from 12 to 19. This choice is driven by an educated guess. The following assumptions were made concerning the LOS.

- The LOS at the CCU is corrected for the additional waiting time (based on measurements).

The additional waiting time was 27% of the original length of stay. The ALOS at the CCU now equals $44 - 0.27*44 = 44 - 12 = 32$ h(1.3 days).

- The LOS at the NC is corrected for the additional time at the CCU. Furthermore a LOS-reduction of 20% is assumed. The ALOS at the NC now equals $0.80*(164+12) = 141$ h(5.9 days).

Table 5 shows the results. The solution area is defined by those values which are closest to 5%.

Table 5 makes clear that several bed combinations are possible to meet the service level requirement. In this case the optimal solution is defined as the one with minimal personnel costs. In order to determine personnel costs for each bed combination the following conversion rates are used:

- 2.2 fte (full time equivalent) per CCU bed
- 0.95 fte per NC bed

In Table 6 the costs are given for each bed combination. The cheapest combination within the solution area is 8 CCU beds and 16 NC beds (32.8 fte). Note that the model allows NC-patients on the CCU. For both hospital professionals and patients this is an undesirable situation as the patient is fit to go to the ward and should be transferred. In this scenario on average 0.28 bed at the CCU is occupied by a NC patient.

The occupancy rate (ρ) at the CCU is now 55%, which means that on average only 4.4 beds out of eight are occupied. The average number of beds occupied at the NC equals 12.5 (ρ=12.5/16=78%). Thus, the amount of reserve capacity which is needed to meet the service level requirement is substantial. As usual a balance between quality and costs must be found.

As mentioned in Section 3.3 the LOS is not a constant of nature. For the optimal solution a sensitivity analysis for the LOS at the CCU has been performed (Fig. 8). A reduction of the LOS with 12 h (0.5 day) reduces the blocking

Table 5 Relation blocking% at CCU and bed distribution over the care chain

CCU Beds	NC Beds							
	12	13	14	15	16	17	18	19
5	*27.2*	*25.8*	*24.9*	*24.3*	*23.95*	*23.76*	*23.66*	*23.61*
6	*20.5*	*18.7*	*17.3*	*16.3*	*15.7*	*15.28*	*15.04*	*14.91*
7	*15.7*	*13.6*	*12*	*10.8*	*9.9*	*9.36*	*8.99*	*8.77*
8	*12*	*9.9*	*8.3*	*7.1*	***6.14***	***5.51***	***5.09***	***4.82***
9	*9.1*	*7.2*	***5.8***	***4.6***	**3.7**	**3.15**	**2.75**	**2.49**
10	*6.85*	***5.2***	**3.95**	**3**	**2.3**	**1.77**	**1.44**	**1.23**
11	***5.05***	**3.7**	**2.7**	**1.9**	**1.4**	**0.99**	**0.74**	**0.58**
12	**3.64**	**2.6**	**1.8**	**1.21**	**0.8**	**0.55**	**0.38**	**0.27**
13	**2.55**	**1.75**	**1.2**	**0.8**	**0.5**	**0.31**	**0.2**	**0.13**
14	**1.74**	**1.15**	**0.7**	**0.5**	**0.3**	**0.17**	**0.1**	**0.06**
15	**1.15**	**0.74**	**0.5**	**0.3**	**0.16**	**0.09**	**0.05**	**0.03**

Bold italics Solution area (Blocking%≈5%), *Italics* Service level too low (Blocking%>5%), *Bold* Service level too high (Blocking%<5%)

Table 6 Relation costs (in fte) and bed distribution

CCU Beds	NC Beds							
	12	13	14	15	16	17	18	19
5	*22.4*	*23.35*	*24.3*	*25.25*	*26.2*	*27.15*	*28.1*	*29.05*
6	*24.6*	*25.55*	*26.5*	*27.45*	*28.4*	*29.35*	*30.3*	*31.25*
7	*26.8*	*27.75*	*28.7*	*29.65*	*30.6*	*31.55*	*32.5*	*33.45*
8	*29*	*29.95*	*30.9*	*31.85*	***32.8***	***33.75***	***34.7***	***35.65***
9	*31.2*	*32.15*	***33.1***	***34.05***	**35**	**35.95**	**36.9**	**37.85**
10	*33.4*	***34.35***	**35.3**	**36.25**	**37.2**	**38.15**	**39.1**	**40.05**
11	***35.6***	**36.55**	**37.5**	**38.45**	**39.4**	**40.35**	**41.3**	**42.25**
12	**37.8**	**38.75**	**39.7**	**40.65**	**41.6**	**42.55**	**43.5**	**44.45**
13	**40**	**40.95**	**41.9**	**42.85**	**43.8**	**44.75**	**45.7**	**46.65**
14	**42.2**	**43.15**	**44.1**	**45.05**	**46**	**46.95**	**47.9**	**48.85**
15	**44.4**	**45.35**	**46.3**	**47.25**	**48.2**	**49.15**	**50.1**	**51.05**

percentage with approximately 5% and is therefore significant. For this reason, quality improvement programs to reduce the ALOS are very useful.

For the optimal solution a sensitivity analysis for the number of arrivals at the CCU is performed. If the number of arrivals per day increases from 3.3 to 3.9 (+18%) the percentage refused admissions doubles to 10%.

When interpreting these results one must be careful as it just describes one particular case. The way in which hospitals deal with the emergency cardiac in-patient flow differs. Besides, the emergency patient flow is a very dynamic care chain. The number of arrivals is time-variant, ALOS-values are not constant and are strongly affected by congestion. As a consequence it is crucial that the emergency care chain is flexible and that the crucial parameters of this patient flow such as arrival rate, LOS and the number of refused admissions are measured on a regular basis.

7 Conclusion

This paper describes how OR techniques have been applied in modeling the emergency cardiac in-patient flow. Bottlenecks at the First Cardiac Aid (FCA), Coronary Care Unit (CCU) and Normal Care clinical ward (NC) have been identified and the impact of variation has been analyzed. The outcome of this study can be split in general and specific results. Some of the general results are not original but a confirmation of earlier results. Due to the impact of process variation on hospital operations and the logistical problems that are still observed in day-to-day practice we chose to mention this outcome again.

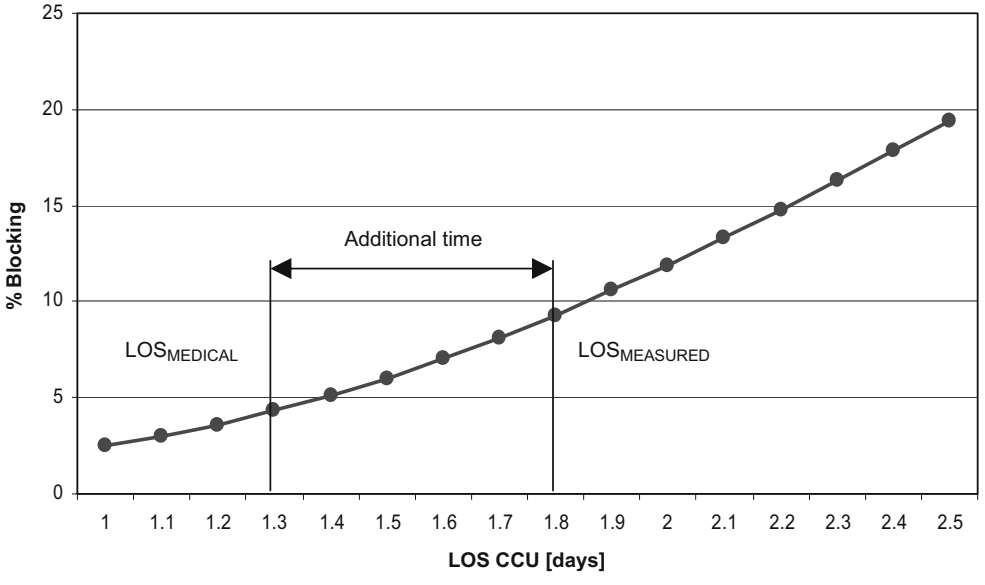

Fig. 8 Sensitivity analysis of the LOS at the CCU

7.1 General results

- A calculation based on average data, regarding the number of arrivals and LOS, does not meet the actual capacity requirements and will most certainly result in frequent operational difficulties. The ultimate consequence is a refused admission. At this moment many hospital professionals, managers and policy makers are not conscious of this flaw of averages.
- The length of stay (LOS) of patients in hospital is highly variable ($C_v \geq 1$) and congestion or chain effects influence the average length of stay (ALOS). This additional waiting time can be as high as 20–30% of the ALOS.
- The characteristics of arrival patterns and LOS distributions result in large workload variations at nursing units.
- The strong focus of hospital management on raising occupancy rates is unrealistic and counterproductive. Larger hospital units can operate at higher occupancy rates than smaller ones while attaining the same percentage of refused admissions (economies of scale). Therefore, one target occupancy rate for all hospital units is not feasible.
- The 85% target is only attainable for units with a minimum of approximately 50 beds. Using the same target for smaller units results in large numbers of refused admissions.
- A small group of patients consumes an enormous and disproportional part of the available resources. In terms of number of patients this group is little but in terms of total resource consumption this group is vital. Nevertheless most of the attention goes out to the larger group. This is known as Pareto's principle.

7.2 Specific results

- Refused admissions at the FCA are primarily caused by unavailability of beds downstream the care chain (CCU and NC).
- The variation in arrival rate at the FCA increases the workload during office hours with 62%.
- For a maximum of 2% refused admissions at the FCA a great amount of reserve capacity is required at the CCU and NC. The number of beds required at the NC and CCU is, respectively, 8 and 16. The occupancy rate for this 'optimal' situation is, respectively, 55 and 78%. For the CCU this means two extra beds are required, a capacity expansion of 2/6=33%.

This paper ends with a discussion and recommendations for further research. An interesting topic is to what extent the long and heavy tail of the LOS distribution, as shown in Fig. 4, can be influenced or shortened. Most studies assume that this feature of LOS is inherent to health care processes. In other words, patients with prolonged hospital stay cannot be denied, we just have to cope with them and offer them the best possible treatment.

As mentioned in Section 3 the group of patients with prolonged hospital stay might be small but their resource consumption is disproportional, thus, they deserve a lot of attention. The introduction of dedicated 'long-stay' meetings where the treatment of these patients is matter of discussion might have a positive effect on the tail of the distribution.

Obviously a high degree of reserve capacity is only one possible solution to decrease the number of refused admissions. It is also a very expensive choice as the beds are not used in an efficient way. Better and more effective is to benefit from the economies of scale. Therefore, merging departments is a good way to increase operational efficiency. A larger department is more flexible and the probability of a refused admission decreases. Hospital professionals seem to be aware of this phenomenon. The last decade more and more specialized intensive care units (cardiac, general, etc) merge into larger general units.

In this paper the impact of variability (in both LOS and arrivals) on capacity requirements has been described. Some of the variability in health care processes, such as the fluctuation in emergency arrivals, is natural and cannot be influenced. Another part of the variation in care chains is introduced by ourselves and is artificial (such as OR-schedules). This non-natural variation must be reduced or eliminated.

However, to reduce the number of refused admissions and consequently raise the quality of care this variability must be taken very seriously. In the present situation a high degree of reserve capacity is required, even though this is not a very economical solution. We have to start thinking about new ways of organizing the emergency patient flow.

References

1. Bagust A et al. (1999) Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. Br Med J 319:155–158
2. Black D, Pearson M (2002) Average length of stay, delayed discharge, and hospital congestion. Br Med J 325:610–611
3. Brennecke R, Kadel C (1995) Requirements for quality assessment in coronary angiograph and angioplasty. Eur Heart J 16:1578–1588
4. Davis JL, Massey WA, Whitt W (1995) Sensitivity to the service-time distribution in the nonstationary Erlang Loss model. Manage Sci 41(6):1107–1116
5. Gallivan S et al. (2002) Booked inpatient admissions and hospital capacity: mathematical modelling study. Br Med J 324:280–282
6. Green LV (2004) Capacity planning and management in hospitals. In: Brandeau ML, Sainfort F, Pierskalla WP (eds) Operations

research and health care. A handbook of methods and applications. Kluwer, London
7. Green LV (2002) How many hospital beds? Inquiry—Blue Cross and Blue Shield Association 39:400–412
8. Green LV, Nguyen V (2001) Strategies for cutting hospital beds: the impact on patient service. Health Serv Res 36:421–442
9. Groothuis S et al. (2004) Predicting capacities required in cardiology units for heart failure patients via simulation. Comput Methods Programs Biomed 74:129–141
10. Harper PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. J Oper Res Soc 53: 11–18
11. Koizumi et al. (2005) Modeling patient flows using a queuing network with blocking. Health Care Manage Sci 8:49–60
12. Laffel G, Blumenthal D (1989) The case for using industrial quality management science in health care organizations. JAMA 262:2869–2873
13. Little JDC (1961) A proof of the queueing formula L=λ W. Oper Res 9:383–387
14. Norris RM (2000) Coronary disease: the natural history of acute myocardial infarction. Heart 83:726–730
15. Ridge JC et al. (1998) Capacity planning for intensive care units. Eur J Oper Res 105:346–355
16. Saunders CE et al. (1989) Modeling emergency department operations using advanced computer simulation models. Ann Emerg Med 18:134–140
17. Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: a survey. J Oper Res Soc 50:109–123
18. Tijms HC (2003) A first course in stochastic models. In: Algorithmic analysis of queues, Chap. 9. Wiley, Chichester
19. Young JP (1965) Stabilization of inpatient bed occupancy through control of admissions. Journal of the American Hospital Association 39:41–48