Ann Oper Res (2010) 178: 23–43
DOI 10.1007/s10479-009-0647-8

Dimensioning hospital wards using the Erlang loss model

A.M. de Bruin · R. Bekker · L. van Zanten · G.M. Koole

Published online: 15 October 2009

Abstract How many beds must be allocated to a specific clinical ward to meet production targets? When budgets get tight, what are the effects of downsizing a nursing unit? These questions are often discussed by medical professionals, hospital consultants, and managers. In these discussions the occupancy rate is of great importance and often used as an input parameter. Most hospitals use the same target occupancy rate for all wards, often 85%. Sometimes an exception is made for critical care and intensive care units. In this paper we demonstrate that this equity assumption is unrealistic and that it might result in an excessive number of refused admissions, particularly for smaller units. Queuing theory is used to quantify this impact. We developed a decision support system, based on the Erlang loss model, which can be used to evaluate the current size of nursing units. We validated this model with hospital data over the years 2004–2006. Finally, we demonstrate the efficiency of merging departments.

Keywords Clinical ward management · Resource allocation · Refused admissions · Bed occupancy · Bed pooling · Queuing models · Erlang loss model

1 Introduction

Most hospitals organize their available beds into nursing units that are used by one or more clinical disciplines (e.g., general surgery, cardiac surgery, cardiology, obstetrics, gynaecology, pediatric, neurology). Over the years other classifications rather than just clinical service appear, such as length of stay (e.g. short stay, medium stay, and long stay units) (Walley

A.M. de Bruin (✉) · L. van Zanten
VU University Medical Center, Division IV, de Boelelaan 1117 (room PK 6X.185), P.O. BOX 7075,
1007 MB Amsterdam, The Netherlands
e-mail: am.debruin@vumc.nl
url: www.vumc.nl/pica

A.M. de Bruin · R. Bekker · G.M. Koole
Faculty of Sciences, Department of Mathematics, VU University, De Boelelaan 1081a (room R-446),
1081 HV Amsterdam, The Netherlands

et al. 2006), level of care (e.g. intensive, medium and normal care units) or urgency (elective, urgent, and emergency care units). We state that the distribution of hospital beds over the different nursing units is to a great extent based on historically obtained rights. A well-founded quantitative approach is often lacking when hospital management decides about the number of beds. Capacity planning issues are primarily driven by available budgets and target occupancy levels instead of service level standards (e.g. percentage of refused admissions, waiting time etc.).

In addition, there are certain data-issues to mention. The registration in Dutch hospitals is based on the following production parameters: number of admissions, day treatments (normal and heavy), nursing days, and number of out-patient visits. The available budget is primarily based on these parameters via contracts with health insurers. Using these parameters for quantitative analysis of patient flow (e.g. resource allocation issues) is delicate and care is required in doing this. In this paper some new concepts and measurements needed to perform a well-founded analysis of in-patient flow are introduced.

The main contribution of this paper is threefold. First, we provide a comprehensive data analysis for 24 clinical wards in a university medical center. We analyze both the number of admissions and the length of stay distribution to obtain insight in the key characteristics of in-patient flow. Moreover, we present occupancy rates for all wards to indicate the bed utilization.

Our second goal is to demonstrate that in-patient flow can be described by a standard queuing model (Erlang loss model). This queuing model may therefore be an important tool in supporting strategic and tactical managerial decisions considering the size of hospital wards. For instance, the model can be used to determine the number of required operational beds and, hence, the corresponding annual budgets.

The third goal is to illustrate the impact of in-patient flow characteristics on ward sizes. The model provides great additional insight in the specific situation at hospital wards where variation in demand is such an important characteristic. It reveals the non-linear relation between the size of a unit, the probability of a refused admission and the occupancy rate. This matter is often not recognized by hospital professionals.

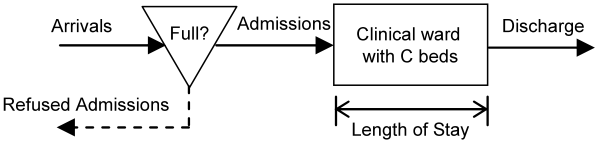
Given the situation where medium and small sized hospital wards operate independently, the model directly underlines the occurrence of operational problems such as unavailability of beds. Moreover, using the model, we illustrate and quantify the potential improvement of operational efficiency by merging clinical wards.

Finally, this queuing model is implemented in a decision support system. This provides hospital management a powerful instrument to evaluate the current size of nursing units and to quantify the impact of bed reallocations and merging of wards.

2 Terminology and definitions

The structural model of the patient flow through a clinical ward is shown in Fig. 1. The different flow parameters are described below. In the next section these parameters are quantified.

Fig. 1 Structural model of patient flow through a clinical ward



Arrivals From the left in Fig. 1 patients arrive to the system and are admitted if there is a free bed available. An arriving patient who finds all beds occupied is refused and leaves the system. We distinguish scheduled and unscheduled arrivals.

Refused admissions If all beds are occupied the patient is 'blocked' and is counted as a refused admission. In practice, a refused admission can result in a diversion to another hospital or an admission to a non-preferable clinical ward. Many wards have to deal with refused admissions due to unavailability of beds. The number of refused admissions can be interpreted as a service level indicator and is important for the quality of care. As this number is hard to measure we do not know how many patients are turned away. In Sect. 4.2 we demonstrate a method to approximate the number of refused admissions.

Admissions Hospitals keep record of the number of admissions. The total amount of admissions is equal to the sum of a number of production parameters such as admissions, day treatments, and transfers (patients coming from another medical discipline). The patients of one medical discipline are often spread over several wards. For example, cardiac patients can be admitted at the normal care clinical ward, the coronary care unit or first cardiac aid. In order to analyze specific wards and evaluate their performance we have to do some basic data manipulation to get the correct flow parameters on the level of nursing units (de Bruin et al. 2007), see Sect. 3.1.

Length of stay The time spent in the ward is entitled length of stay, often abbreviated as LOS, after which the patient is discharged or transferred to another ward. The LOS is easily derived from the hospital information system as both time of admission and time of discharge are logged on individual patient level. Length of stay is affected by congestion and delay in the care chain (Koizumi et al. 2005), which means that careful interpretation is required, see Sect. 3.2.

Beds The capacity of a ward is measured in terms of operational beds. The number of operational beds, a management decision, is used to determine the available personnel budget. This is done via a staffing ratio per operational bed (e.g. 1 full time equivalent (fte) per normal care bed). The number of operational beds is generally fixed and evaluated on a yearly basis.

From day to day, the actual number of open or staffed beds slightly fluctuates (due to illness, holiday and patient demand) (Green et al. 2007). Note that the number of physical bed positions is not necessarily equal to the number of operational beds; for most wards the number of physical beds is larger.

3 Data analysis

Computerized records of all admissions to 24 clinical wards in a university medical center, both normal care (NC) and intensive care units (ICU), have been analyzed over the years 2004–2006. These data were used to quantify the number of admissions (Sect. 3.1), length of stay (Sect. 3.2), and occupancy rate (Sect. 3.3). Due to the specific ward characteristics we did not include the emergency department (ED), first cardiac aid (FCA), and short stay unit (SSU) in this study. Table 1 gives an overview of the wards included and the corresponding number of operational beds.

Table 1 Wards included in this study and number of operational beds

Ward description	Operational beds (2004)	Operational beds (2005)	Operational beds (2006)
Coronary Care Unit (CCU)	6	6	6
Intensive Care Unit surgical	14	14	14
Intensive Care Unit medical	12	13	14
Pediatric Intensive Care Unit (PICU)	9	9	9
Neonatal Intensive Care Unit (NICU)	20	19	15
Medium Care	7	7	9
Special Care cardiac surgery	6	6	6
NC Cardiac surgery and cardiology	28	28	28
NC Gynaecology	37	37	37
NC Hematology	21	21	21
NC Surgical oncology	27	27	27
NC Internal medicine unit 1	20	20	20
NC Internal medicine unit 2	20	20	20
NC Pediatric unit 1	26	26	26
NC Pediatric unit 2	23	23	25
NC Otolaryngology (ear/nose/throat)	29	26	25
NC Internal lung	23	23	23
NC Neurosurgery and orthopedic surgery	27	27	30
NC Neurology	31	26	24
NC Obstetrics	42	37	31
NC Internal oncology	26	27	27
NC Ophthalmology	21	15	14
NC Trauma surgery	32	30	33
NC Vascular surgery	21	18	23

3.1 Admissions

In this subsection the admissions are analyzed. First, in Sect. 3.1.1 the scheduled (elective) admissions are quantified and in Sect. 3.2.2 the unscheduled (urgent, emergent) admissions are described. For each ward the admission pattern is compared with the Poisson distribution. Many arrival processes, especially unscheduled, have been shown to be well approximated by a Poisson process (Young 1965).

Table 2 summarizes the number of annual admissions. In 2006 approximately 45% of all admissions were unscheduled. This fraction of unscheduled admissions ranges from 7% for hematology to 84% for obstetrics.

3.1.1 Scheduled admissions

The total number of scheduled (or elective) admissions over the years 2004–2006 was respectively 17969, 17101, and 18001. For each ward we constructed histograms with the daily number of scheduled admissions on the horizontal axis and the frequency (number of days in a year that this number of daily admissions occurred) on the vertical axis. Figure 2

Table 2 Hospital admissions (2004–2006)

Ward description	2004		2005		2006	
	Sched.	Unsched.	Sched.	Unsched.	Sched.	Unsched.
Coronary Care Unit	292	832	221	772	207	741
Intensive Care Unit surgical	419	325	341	373	330	392
Intensive Care Unit medical	354	325	341	338	386	328
Pediatric Intensive Care Unit	139	268	100	296	103	245
Neonatal Intensive Care Unit	143	278	88	429	88	384
Medium Care	257	316	387	430	491	475
Special Care cardiac surgery	572	112	503	105	496	119
NC Cardiac surgery and cardiology	1384	618	1321	664	1280	620
NC Gynaecology	1275	1203	1051	1242	1248	1248
NC Hematology	1707	89	2003	121	2178	160
NC Surgical oncology	803	353	620	343	776	434
NC Internal medicine unit 1	298	691	204	725	211	721
NC Internal medicine unit 2	596	712	668	725	665	607
NC Pediatric unit 1	944	566	923	703	996	641
NC Pediatric unit 2	953	742	844	789	788	659
NC Otolaryngology (ENT)	834	340	1115	455	1064	305
NC Internal lung	875	420	825	401	781	411
NC Neuro- and orthopedic surgery	885	391	993	389	1061	457
NC Neurology	365	838	233	942	252	903
NC Obstetrics	1288	2708	749	3575	658	3406
NC Internal oncology	1132	305	1230	355	1007	388
NC Ophthalmology	1206	227	1255	216	1636	217
NC Trauma surgery	607	630	494	636	565	821
NC Vascular surgery	641	289	592	328	734	314
Total	17969	13578	17101	15352	18001	14996

gives an example for the normal care hematology, where the observed admission pattern is compared with the Poisson distribution.

The average number of scheduled arrivals per day (parameter λ) equals 4.664. Clearly, the Poisson distribution does not give a good fit. A possible explanation is that elective patients are generally admitted during weekdays (Mon–Fri) and hardly in the weekend (Sat–Sun). This explains the peak bar at “0”, corresponding with the weekend days. Therefore we split the scheduled admissions in weekdays and weekends. See Fig. 3 for the results. Since the average number of arrivals per day during the weekend is very small for most clinical wards, we focus here on the number of arrivals during weekdays.

For a Poisson random variable, the mean and variance are equal. As a first quantitative indication for the variability in the number of arrivals, we determined the ratio of the variance in the number of arrivals and the average number of arrivals per weekday (this ratio is 1 for a Poisson random variable).

For hematology, this variance/mean ratio equals $6.435/5.764 = 0.896$. Using the data of 2005, the variance/mean ratio ranges from 0.658 to 1.759 for the 24 wards, indicating at least that the number of scheduled arrivals is highly variable. This fact is rather remarkable

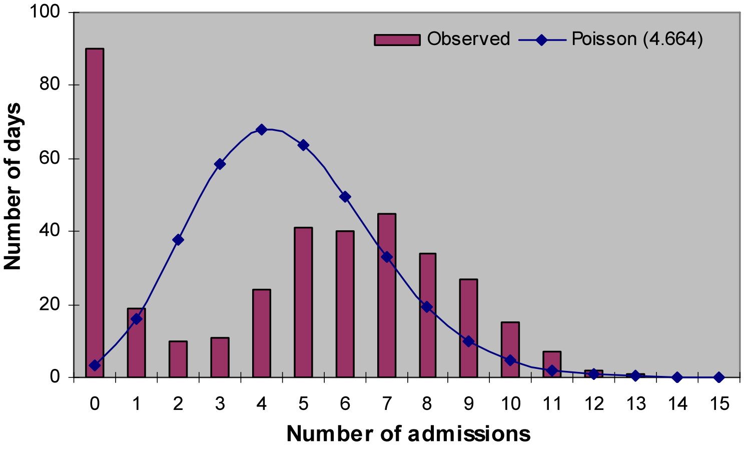


Fig. 2 Distribution of scheduled admissions for hematology (2004)

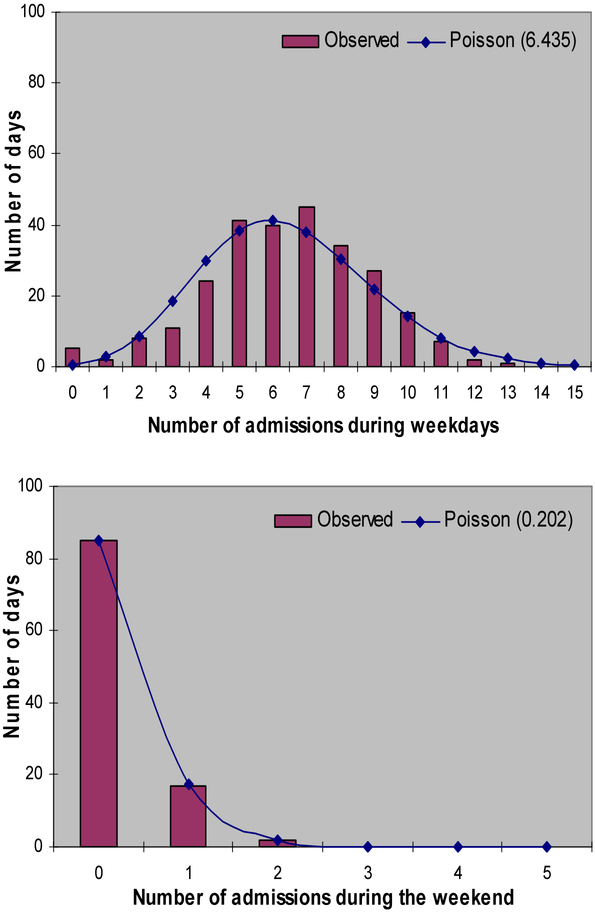


Fig. 3 Distribution of scheduled admissions split in weekdays and weekend for hematology (2004)

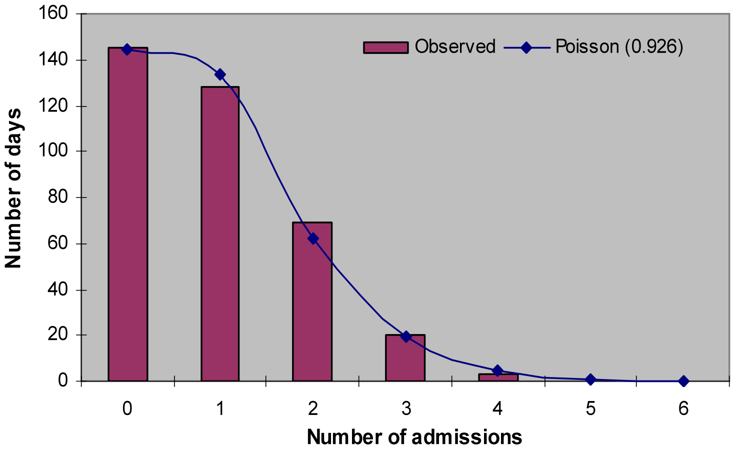


Fig. 4 Distribution of unscheduled admissions for the ICU medical (2005)

and also reported in other studies (Walley et al. 2006). Most professionals would expect the elective patient flow to be more steady.

In Appendix A some results can be found on formal statistical tests to assess the goodness of fit. The general conclusion is that for roughly half of the 24 wards the scheduled admissions, if separated in weekdays and weekends, are described well by a Poisson distribution. We like to stress that for practical purposes it is not required that the number of admissions exactly follows the laws of the Poisson distribution. The key point for practical modeling purposes is that the variability in the number of admissions is generally well captured by the Poisson distribution, making this a reasonable assumption for the remaining wards as well. This is also exemplified by the stationary peakedness approximation for $G/G/s/s$ queues, where the variability in the arrival process is approximated using the variance to mean ratio (Whitt 1984).

An interesting article (Gallivan 2008), recently published, challenges the view that modeling should necessarily be subject to formulaic calibration, validation and sensitivity analysis to establish ‘accuracy’. In this paper the author recalls the quote, “All models are wrong, but some are useful”, of George Box, an eminent statistician. We believe that the model presented in Sect. 4 provides valuable insight when determining the number of beds required at a specific hospital ward. In their role as hospital consultants, the authors have broad experience in advising management on capacity decisions. The model, although not formally validated, has been proven useful for both hospital professionals and management and contributes to better decision making.

3.1.2 Unscheduled admissions

The total number of unscheduled (urgent and emergent) admissions over the years 2004–2006 was respectively 13578, 15352, and 14996. In studies of unscheduled admissions the assumption of a Poisson arrival process has been shown to be realistic (Young 1965).

We constructed the same histograms as for the scheduled arrivals, thus the number of daily admissions is plotted on the x-axis and the frequency (number of days) on the y-axis. Figure 4 gives an example for the medical intensive care unit (2005).

The statistical tests also show that the Poisson distribution provides a good fit, see Appendix A. Intuitively, it is obvious that unscheduled arrivals are independent, because one emergency admission does not have any effect on the next one. Therefore, the assumptions for a Poisson process seem realistic.

3.2 Length of stay

In this section the length of stay (LOS) is examined. First, in Sect. 3.2.1, some key statistics such as the average length of stay (ALOS) and the coefficient of variation are specified for each ward. In Sect. 3.2.2 the LOS distribution is characterized by Lorenz curves and the related Gini-coefficient is introduced. This representation is deliberately chosen as we find it supportive in visualizing the variation in LOS which is an important characteristic of in-patient flow.

3.2.1 Basic LOS statistics

In Table 3 the basic LOS statistics are summarized. In 2006 the ALOS over 24 wards was approximately 4 days, ranging from 1.5 days (obstetrics) to 7.8 days (neonatal intensive care unit). This average is calculated over all admitted patients, thus including day treatments. The coefficient of variation, defined as the ratio of the standard deviation to the mean, is greater than 1 for all wards, except for the special care cardiac surgery. This shows that the LOS at clinical wards is highly variable. This result is also found by Green and Nguyen (2001). The Gini-coefficient will be introduced below.

3.2.2 Lorenz curves

The Lorenz curve is a graphical representation of the cumulative distribution function of a probability distribution. This concept was introduced to represent the concentration of wealth (Lorenz 1905). We use it to illustrate that those patients with prolonged hospital stay take a disproportional part of the available resources. The percentage of patients is plotted on the x-axis, the percentage of resource consumption (in terms of hospitalized days) on the y-axis. Figure 5 gives the Lorenz curves with the lowest and highest Gini-coefficient for the year 2006, respectively the special care cardiac surgery and normal care hematology.

The Gini-coefficient (denoted as G) is a measure of the dispersion of the Lorenz curve (Gini 1912). It is defined as a ratio with values between 0 and 1; the numerator is the area between the Lorenz curve of the distribution and the uniform distribution line; the denominator is the area under the uniform distribution line. Thus, a low Gini-coefficient indicates that the variability in LOS is low, while a high Gini-coefficient indicates a more variable distribution. The following formula was used to calculate the Gini-coefficient for each clinical ward,

$$G = \frac{1}{n}\left(n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right)$$

with

n number of admitted patients to a ward
y_i the observed LOS values in ascending order, where $y_i \leq y_{i+1}$, $i = 1, \dots, n-1$

In Table 3 all Gini-coefficients (G) are specified. For 2006, the minimum is 0.343 for the special care cardiac surgery and the maximum is 0.805 for the NC hematology. Figure 5 clearly shows the different shape of the two curves. Lorenz curves are very useful in identifying those patients with prolonged hospital stay and their disproportional resource consumption. It also illustrates the difference in LOS-characteristics between clinical wards. We note that in general a model based on fixed lengths of stay is not capable of describing the complexity and dynamics of in-patient flow and gives misleading results, also known as the flaw of averages (de Bruin et al. 2007).

Table 3 LOS statistics (2004–2006)

Ward description	2004			2005			2006		
	ALOS [days]	C_V [σ/μ]	Gini [G]	ALOS [days]	C_V [σ/μ]	Gini [G]	ALOS [days]	C_V [σ/μ]	Gini [G]
Coronary Care Unit	1.573	1.692	0.638	1.748	1.311	0.564	1.694	1.313	0.569
Intensive Care Unit surgical	5.676	2.690	0.750	6.021	2.012	0.719	5.397	1.730	0.684
Intensive Care Unit medical	4.957	1.851	0.687	5.417	2.025	0.705	5.147	2.018	0.684
Pediatric Intensive Care Unit	3.533	1.997	0.644	4.168	1.571	0.636	4.180	1.512	0.606
Neonatal Intensive Care Unit	9.025	1.721	0.668	8.335	1.985	0.697	7.778	1.644	0.680
Medium Care	2.211	1.478	0.543	2.619	1.738	0.606	2.374	1.643	0.584
Special Care cardiac surgery	1.524	0.733	0.319	1.668	0.644	0.309	1.715	0.821	0.343
NC Cardiac surgery and cardiology	4.005	1.519	0.567	4.054	1.502	0.570	4.347	1.430	0.570
NC Gynaecology	3.444	1.290	0.552	3.615	1.420	0.570	3.172	1.391	0.558
NC Hematology	3.732	2.667	0.819	3.023	2.591	0.812	2.763	2.651	0.805
NC Surgical oncology	6.571	1.110	0.500	7.374	1.111	0.497	6.448	1.167	0.510
NC Internal medicine unit 1	6.493	1.273	0.579	7.266	1.292	0.553	6.354	1.089	0.528
NC Internal medicine unit 2	4.853	1.469	0.653	4.810	1.713	0.683	4.852	1.627	0.685
NC Pediatric unit 1	3.850	1.449	0.588	3.758	1.470	0.591	3.440	1.486	0.579
NC Pediatric unit 2	3.432	1.434	0.599	4.119	2.243	0.633	4.131	1.750	0.630
NC Otolaryngology (ENT)	5.333	1.552	0.612	4.052	1.544	0.577	4.362	1.649	0.596
NC Internal lung	4.765	1.139	0.520	4.517	1.245	0.541	4.633	1.198	0.541
NC Neuro- and orthopedic surgery	6.340	1.527	0.579	5.441	1.472	0.545	5.256	1.302	0.537
NC Neurology	5.921	1.503	0.609	5.597	1.300	0.587	5.533	1.401	0.590
NC Obstetrics	1.589	1.784	0.688	1.438	2.005	0.707	1.501	2.039	0.698
NC Internal oncology	4.904	1.373	0.572	4.061	1.346	0.574	4.527	1.314	0.562
NC Ophthalmology	2.783	1.150	0.499	2.180	1.225	0.487	1.583	1.186	0.534
NC Trauma surgery	7.695	1.299	0.541	7.641	1.251	0.537	6.833	1.175	0.526
NC Vascular surgery	6.195	1.345	0.554	6.844	1.321	0.550	6.487	1.638	0.583
Average	4.199	1.544	0.595	4.060	1.556	0.594	3.918	1.507	0.591

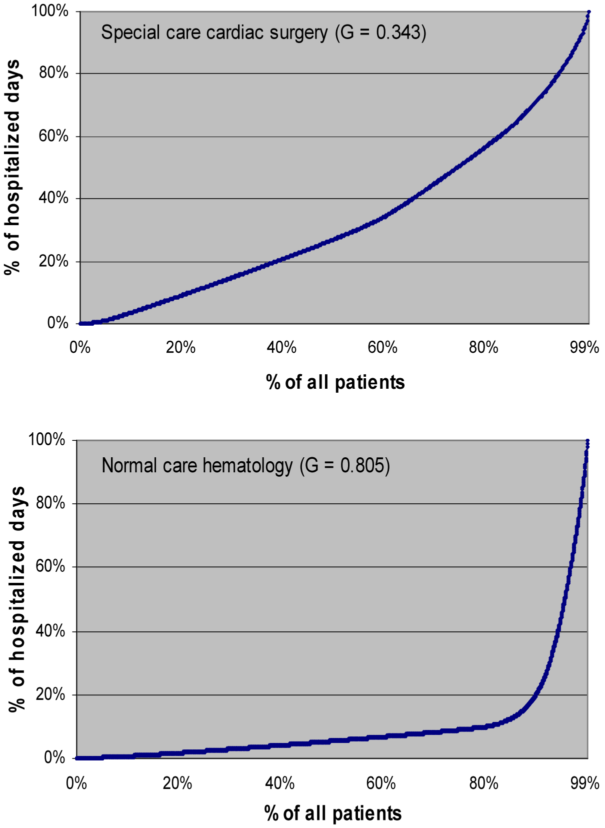


Fig. 5 Lorenz curves for the special care cardiac surgery and NC hematology (2006)

3.3 Occupancy rate

Finally we determined the occupancy rate per ward over the years 2004–2006. Using Little's formula (Little 1961) the occupancy rate is defined as,

$$\begin{aligned} Occupancy &= \frac{Average\ number\ of\ occupied\ beds}{Number\ of\ operational\ beds} \\ &= \frac{Admissions\ (per\ time\ unit) \times ALOS\ (time\ unit)}{Number\ of\ operational\ beds} \end{aligned}$$

See Table 4 for the results per clinical ward. For 2006 the occupancy rate differs from 44% (Pediatric Intensive Care Unit) to 85% (Normal Care internal medicine, unit 2).

Note that this definition of occupancy, which is very common from the perspective of operations research and management science, is *not* used in most Dutch hospitals. The current national definition is based on 'hospitalized days', which is an administrative financial parameter (de Bruin et al. 2007). Under the latter definition occupancy rates greater than 100% are possible which makes the discussion confusing. In our opinion, the definition presented above gives the best insight in the actual utilization of the available capacity. Therefore, this definition is used in this study.

Table 4 Occupancy rate per clinical ward (2004–2006)

Occupancy rates			
Ward description	2004	2005	2006
Coronary Care Unit	80.7%	79.3%	73.3%
Intensive Care Unit surgical	82.6%	84.1%	76.3%
Intensive Care Unit medical	76.8%	77.5%	71.9%
Pediatric Intensive Care Unit	43.8%	50.2%	44.3%
Neonatal Intensive Care Unit	52.0%	62.1%	67.1%
Medium Care	76.6%	86.8%	69.8%
Special Care cardiac surgery	47.6%	46.3%	48.2%
NC Cardiac surgery and cardiology	78.5%	78.7%	80.8%
NC Gynaecology	63.2%	61.4%	58.6%
NC Hematology	87.4%	83.8%	84.3%
NC Surgical oncology	77.1%	72.1%	79.2%
NC Internal medicine unit 1	88.0%	92.5%	81.1%
NC Internal medicine unit 2	87.0%	91.8%	84.5%
NC Pediatric unit 1	61.3%	64.4%	59.3%
NC Pediatric unit 2	69.3%	80.1%	65.5%
NC Otolaryngology (ENT)	59.1%	67.0%	65.4%
NC Internal lung	73.5%	66.0%	65.8%
NC Neuro- and orthopedic surgery	82.1%	76.3%	72.9%
NC Neurology	63.0%	69.3%	73.0%
NC Obstetrics	41.4%	46.0%	53.9%
NC Internal oncology	74.3%	65.3%	64.1%
NC Ophthalmology	52.0%	58.6%	57.4%
NC Trauma surgery	81.5%	78.9%	78.6%
NC Vascular surgery	75.2%	95.8%	81.0%

3.4 Discussion data analysis

In this section we performed an analysis of patient flow characteristics of 24 clinical wards in a university medical center. From the data analysis we found that the number of unscheduled arrivals can be well described by a Poisson distribution. For roughly half of the wards, the Poisson distribution also provides a good fit for scheduled arrivals in case these are split between weekdays and weekends. Given the variability in the number of scheduled arrivals for the remaining half of the wards, the Poisson distribution is a very reasonable simplifying assumption for these wards as well, capturing the apparent fluctuation in the number of scheduled arrivals.

In the sequel, we assume that arrivals occur according to a homogeneous Poisson process and determine the arrival rate using the average number of arrivals per day. In the present paper, we intentionally choose not to model the time-dependent arrival pattern, mainly for practical purposes such as simplicity. A time-dependent model is too refined for our main objective, which is supporting strategic and tactical managerial decisions on sizing hospital wards. A time-dependent model may be primarily used to exploit structural differences in bed occupancy across the week, which is important on the operational level, such as nurse rostering. We note that ignoring this time-dependent arrival pattern leads to a slight underestimation of the average number of required beds. Nevertheless, this impact on the required number of operational beds, and thus personnel budget, is rather limited in most cases.

In other papers these time-varying aspects within service systems are described and its impact (for example on setting staffing requirements) is further quantified (Davis et al. 1995; Green et al. 2007). The impact of a time-varying arrival rate on refused admissions, specifically for in-patient flow, is also described in literature (Bekker and de Bruin 2009).

A second main reason for ignoring the difference between weekdays and weekends is due to the simplicity of use for analysis and implementation. A relatively easy model is effective in developing insight between the size of clinical ward and the feasibility of occupancy rates in relation to the probability of a refused admission, which is another main goal of this study.

4 Mathematical model

In this section the mathematical (queuing) model is described. Recall that we introduced the incidence of refused admissions in our structural model (Fig. 1). Thus, we assume that an arriving patient who finds all beds occupied is blocked and leaves the system. In practice, a refused admission can result in a diversion to another surrounding hospital or an admission to a non-preferable clinical ward within the same hospital.

There are also studies in which delay models, such as the $M/M/s$ queuing model, are applied for capacity planning problems in hospitals (Green 2002; Green and Nguyen 2001). Arriving patients enter a queue when all beds are occupied. The probability of delay is an important performance measure is this type of analysis. Other studies apply an infinite server approach (Gallivan et al. 2002). In this approach the main outcome parameter is the probability of overload when the number of required beds exceeds the number of operational beds. Due to our experience in a university hospital, where diverting patients is a serious issue, we chose to incorporate blocking. For that reason, we applied the $M/G/c/c$ (or Erlang loss) queuing model. To our knowledge, this is the first time that the Erlang loss model is applied in this context.

4.1 The $M/G/c/c$ queuing model

In the $M/G/c/c$ model patients arrive according to a Poisson process with parameter λ. The LOS of an arriving patient is independent and identically distributed with expectation μ. We note that the average service time is often defined as $1/\mu^*$ where μ^* is the service rate in case of exponentially distributed service times. The number of operational beds is equal to c. There is no waiting area, which means that an arriving patient who finds all beds occupied is blocked. The fraction of patients which is blocked can be calculated with the following formula,

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

Note that this particular model is insensitive for the LOS-distribution and is valid for general service times. The occupancy rate is defined as,

$$\mathit{Occupancy\ rate} = \frac{(1-P_c)\lambda\cdot\mu}{c} \tag{ii}$$

This is equivalent to the expression at the beginning of Sect. 3.3. The term $\lambda{\cdot}\mu$ is often referred to as the offered load to the system.

4.2 Approximating the number of arrivals

In order to apply the Erlang loss model and for purposes of validation we need to quantify the number of arrivals (λ). As the hospital information system only registers the number of admissions and the number of refused admissions is generally unknown we have to approximate λ. This can be accomplished by using expressions (i) and (ii). First, the numerator in (ii) is equal to the average number of occupied beds,

$$\textit{Average number of occupied beds} = \lambda \cdot \mu(1 - P_c)$$

Then, after substitution of P_c using (i), we get the following expression,

$$\textit{Average number of occupied beds} = \lambda \cdot \mu \left(1 - \frac{(\lambda\mu)^c/c!}{\sum_{k=0}^{c}(\lambda\mu)^k/k!}\right) \tag{iii}$$

The average number of occupied beds can be obtained from the hospital information system. Furthermore, the number of beds (c) and the ALOS (μ) are known variables. After substitution in (iii) λ is the only parameter left unknown. Finally, the number of arrivals is determined numerically for all clinical wards.

5 Validation

In the previous section we approximated λ using the Erlang loss equation. The assumptions for this model regarding the arrival process (Poisson) and the LOS distribution (general) are discussed in Sect. 3. One of the main objectives of this study is to approximate the number of *required* beds at a clinical ward. Therefore, we first have to determine the distribution of the number of *occupied* beds. In Fig. 6 the observed number of occupied beds (at 08.00 am) is compared with the Erlang loss model, as an example, for three wards. The loss model curve has been curtailed at the number of operational beds which was respectively 26 (Neurology), 6 (CCU) and 14 (Ophthalmology).

Figure 6 reveals that on some days the number of occupied beds exceeds the number of operational beds. This means that when patient demand is high, ward management can decide to temporarily open an extra few beds. Of course this puts the available staff under great working pressure and is only possible for wards where the number of physical bed positions exceeds the number of operational beds, also see Sect. 2.

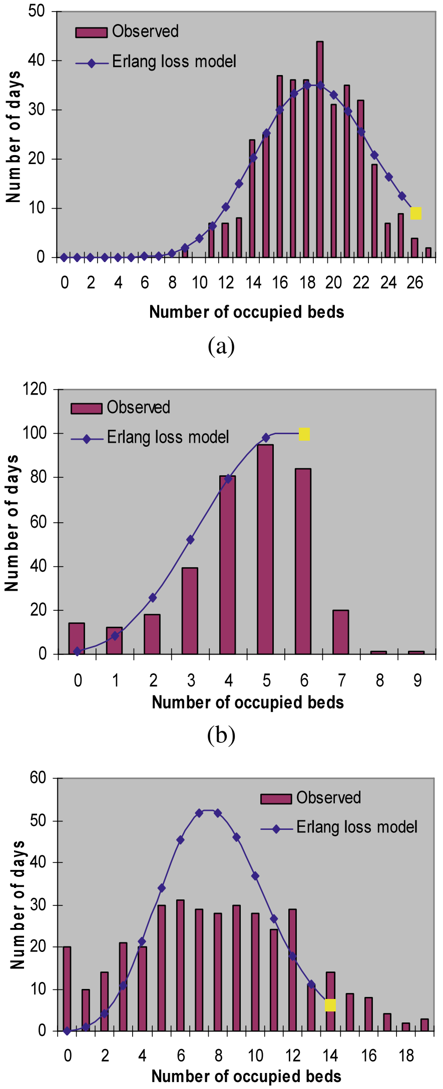
Furthermore, the model seems reasonable for the NC Neurology and the CCU but is fairly poor for the NC Ophthalmology. These figures were created for all 24 wards to identify how good the Erlang loss model fits the observed number of occupied beds. To quantify the goodness of fit we introduced a validation measure. First we define,

- $P_i =$ probability that i beds are occupied in the Erlang loss model
- $Preal_i =$ probability that i beds are occupied in reality
- $D_i = Preal_i - P_i$ for $i = 0, 1, \ldots, c-1$
- $D_c = \sum_{k=c}^{\infty} Preal_k - P_c$

The final formula above is used to compare the possible different situations with a fully occupied ward. Remind that in practice the number of occupied beds may occasionally exceed the number of operational beds, whereas this is not possible in the Erlang loss model.

To compare to what extent the empirical distribution and the bed occupancy distribution given by the Erlang loss model are similar, we define our performance measure for the

Fig. 6 (**a**) The observed number of beds occupied versus the Erlang loss model, NC Neurology (2005). (**b**) The observed number of beds occupied versus the Erlang loss model, Coronary Care Unit (CCU) (2005). (**c**) The observed number of beds occupied versus the Erlang loss model, NC Ophthalmology (2005)



(a)

(b)

(c)

goodness of fit as the sum of the absolute differences between the two probabilities:

$$Goodness\ of\ fit = 1 - \frac{1}{2}\sum_{i=0}^{c} |D_i|$$

In terms of distribution functions, our measure can thus be interpreted as the amount of probability mass that the empirical and Erlang loss based distribution have in common. Therefore, the measure is a number between 0 and 1, where 0 indicates a very poor fit (no probability mass in common) and 1 means the probabilities are equal for all number of occupied beds (exact match between distribution functions). We note that this measure for

Table 5 Goodness of fit of the Erlang loss model describing the number of occupied beds

Ward description	Goodness of fit	Ward description	Goodness of fit
Coronary Care Unit (CCU)	0.941	NC Internal medicine unit 2	0.935
Intensive Care Unit surgical	0.854	NC Pediatric unit 1	0.841
Intensive Care Unit medical	0.756	NC Pediatric unit 2	0.926
Pediatric Intensive Care Unit (PICU)	0.860	NC Otolaryngology (ear, nose, throat)	0.889
Neonatal Intensive Care Unit (NICU)	0.891	NC Internal lung	0.921
Medium Care	0.838	NC Neurosurgery and orthopedic surgery	0.872
Special Care cardiac surgery	0.856	NC Neurology	0.902
NC Cardiac surgery and cardiology	0.833	NC Obstetrics	0.945
NC Gynaecology	0.874	NC Internal oncology	0.888
NC Hematology	0.658	NC Ophthalmology	0.768
NC Surgical oncology	0.759	NC Trauma surgery	0.835
NC Internal medicine unit 1	0.925	NC Vascular surgery	0.815

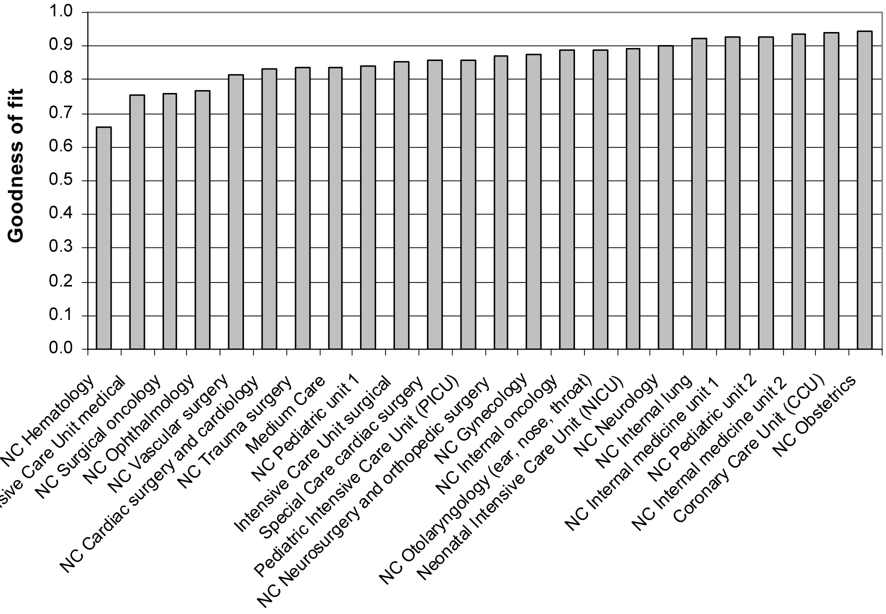


Fig. 7 Goodness of fit of the Erlang loss model for all 24 wards

goodness of fit is inspired on the average absolute prediction error (Kleijnen et al. 2000). The outcome of this measurement for all 24 wards is presented in Table 5.

The average goodness of fit is 0.86 meaning that the empirical and model-based bed-occupancy distributions have on average 86% of the probability mass in common. For most wards, the model describes the number of occupied beds very well, especially those with a high percentage of unscheduled admissions (Obstetrics, CCU and internal medicine). However, for a couple of wards the fit is rather poor (NC hematology and Ophthalmology). See Fig. 7 for a general overview.

Another important and more general observation is that the variability in both the number of arrivals per day and LOS results in large workload fluctuations, i.e. a considerable variability in bed occupancy. Therefore, the occurrence of both over and underutilization is inevitable and flexibility in workforce planning at clinical wards is crucial.

6 Decision support system

One of the objectives of this study was to develop a decision support system (DSS) upon which hospital management can make well founded decisions about hospital ward size. The model has to be user-friendly and technologically accessible. Encouraged by the validation of the Erlang loss model in the previous section we implemented this queuing model in MS Excel using Visual Basic for Applications (VBA). Historical data of admissions, ALOS and number of operational beds are automatically imported from the hospital information system to calculate the occupancy and to approximate the number of arrivals. The tool can

Table 6 Number of required beds for different levels of blocking in current situation (2006)

Ward description	Operational beds (2006)	Number of daily arrivals (λ)	Number of beds required for: 2% Blocking	5% Blocking	10% Blocking
Coronary Care Unit	6	3.52	12	10	9
Intensive Care Unit surgical	14	2.26	19	17	15
Intensive Care Unit medical	14	2.14	18	16	14
Pediatric Intensive Care Unit	9	0.97	9	8	7
Neonatal Intensive Care Unit	15	1.36	17	15	14
Medium Care	9	3.06	13	12	10
Special Care cardiac surgery	6	1.79	8	7	6
NC Cardiac surgery and cardiology	28	5.62	33	30	27
NC Gynecology	37	6.84	30	27	24
NC Hematology	21	7.57	29	26	24
NC Surgical oncology	27	3.55	32	29	26
NC Internal medicine unit 1	20	2.9	27	24	21
NC Internal medicine unit 2	20	4.17	29	26	23
NC Pediatric unit 1	26	4.5	23	21	18
NC Pediatric unit 2	25	4.02	24	22	20
NC Otolaryngology (ENT)	25	3.8	24	22	19
NC Internal lung	23	3.32	23	21	18
NC Neuro- and orthopedic surgery	30	4.26	31	28	25
NC Neurology	24	3.29	26	24	21
NC Obstetrics	31	11.14	25	22	20
NC Internal oncology	27	3.86	25	23	20
NC Ophthalmology	14	5.18	14	13	11
NC Trauma surgery	33	3.97	36	33	30
NC Vascular surgery	23	3.19	29	26	23
Total	507	95.98	556	502	445

be used for both evaluating the current size of nursing units and determining the impact of merging departments. This last feature is important as the potential increase in productivity is substantial if economies of scale are applied properly.

6.1 Evaluating the current size of clinical wards

First, the current size of the hospital wards is evaluated. For each ward the number of arrivals is determined as described in Sect. 4.2. Then the number of required beds is computed for three levels of blocking (2, 5, and 10%). See Table 6 for the results.

Which percentage of blocking (or refused admissions) is reasonable and thus acceptable is subject to discussion. Hospital policy makers often refer to a 5% target but the consequence of such a choice in terms of capacity requirements is very often not recognized. Table 6 reveals that for a 5% target the total number of required beds over 24 wards is approximately 500 which is not far off the number of operational beds in 2006, which is 507. In Fig. 8 the number of operational beds is compared with the number of required beds for each ward (5% blocking).

The ratio (required beds/operational beds) varies from 0.71 to 1.67. In other words, some wards have too many beds while others have a serious shortage. This unequal allocation of beds is also reflected in the range of occupancy rates (Table 4) and is subject to discussion by hospital professionals and management.

6.2 The impact of merging departments on efficiency

In most queuing systems economies of scale occur, which means that larger service systems can operate at higher occupancy rates than smaller ones while attaining the same percentage

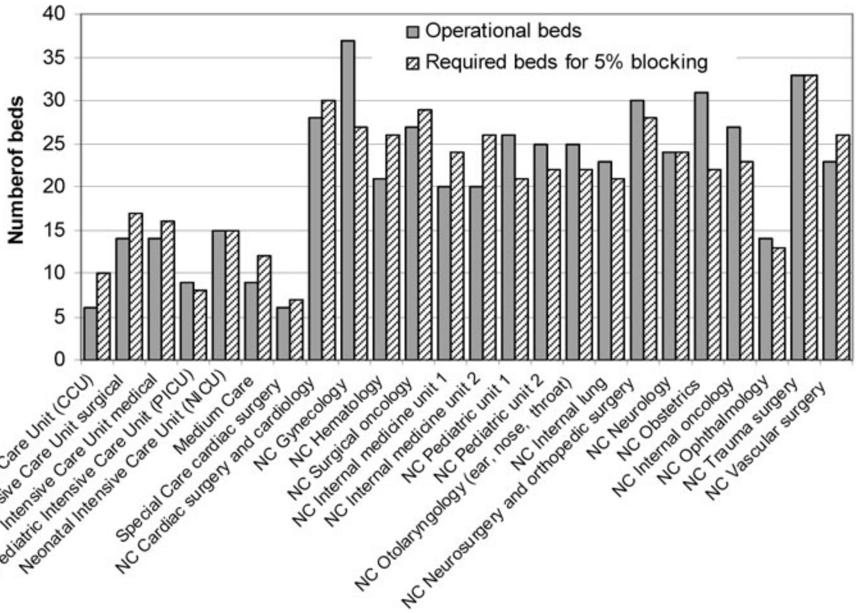


Fig. 8 Number of required beds versus the number of operational beds

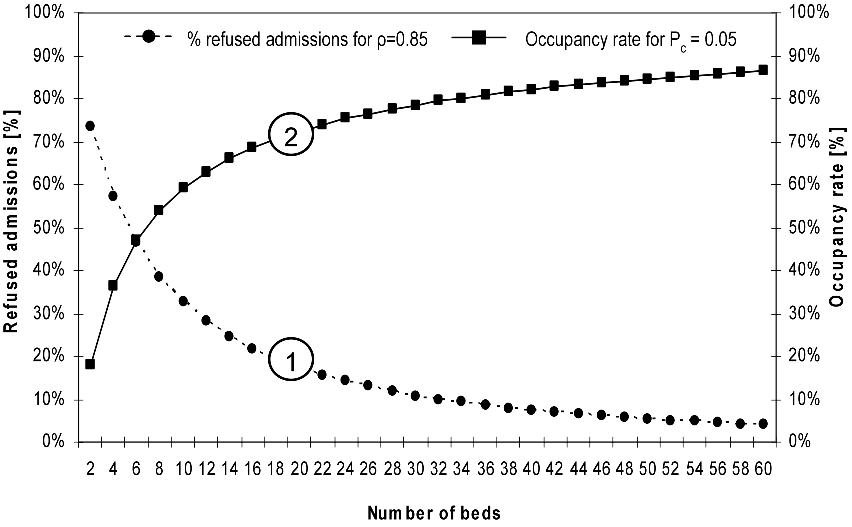


Fig. 9 Relation between number of beds, fraction of refused admissions (P_c), and occupancy rates (ρ)

of blocking or delay (Whitt 1992). Figure 9 (de Bruin et al. 2007) illustrates the dramatic impact for the situation in most Dutch hospitals where ward sizes are relatively small and dispersed and where the 85% target occupancy rate has developed into a golden standard. In Fig. 9 two graphs are shown for varying ward sizes ($2 \le c \le 60$):

1. The percentage of refused admissions given that the occupancy rate (ρ) equals 85%
2. The occupancy rate for a maximum of 5% refused admissions

Especially smaller units, such as the CCU (6 operational beds), have trouble keeping a bed available for an arriving patient. If the probability of a refused admission has to be kept low (for example 5%) the occupancy rate drops below 50% (see Fig. 9, graph 2). Off course this is not a very economical way to use a scarce and expensive resource such as hospital beds. Li et al. studied the relation between profits and occupancy rates in more detail using a multi-objective bed allocation model (Li et al. 2008). If economies of scale are applied properly, by merging departments (bed pooling) or mixing patient flows, both an acceptable service level (in terms of refused admissions) and an economical viable occupancy rate can be realized.

6.3 Case study

In this subsection an example is described in which the DSS can be used to explore the potential benefit of merging departments. We do this by virtually merging the coronary care unit (CCU), medium care (MC) and the special care cardiac surgery (SC) because the level and type of care is similar. By selecting different units the user can experiment easily and fast with different scenarios in the DSS. Table 7 summarizes the statistics for 2006. The results of the scenario where these three units are merged is also presented in this table.

In the current situation these three units have 21 beds together and the fraction of refused admissions ranges from 5.61% (SC) to 26.2% for the CCU. The total number of required beds, for 5% refused admissions at each ward, is 29.

Table 7 The effect of merging departments on operational efficiency: case study

Parameters (2006)	CCU	MC	SC
Operational beds	6	9	6
ALOS [days]	1.694	2.374	1.715
Occupancy rate	73.3%	69.8%	48.2%
Fraction of refused admissions*	26.2%	13.5%	5.61%
Number of beds required for 5% blocking*	10	12	7
After merging CCU, MC, and SC			
ALOS (weighted)		1.96	
Number of beds required for 5% blocking*		22	
Occupancy rate		71.7%	

*Calculated with the Erlang loss model

After the merge the total number of required beds for this same service level (5% blocking) is 22, just one more bed than the number of beds at this moment. Thus, the improvement in operational efficiency is significant, and can be attained by just creating a larger scale.

7 Conclusion

In this study we applied the Erlang loss (or $M/G/c/c$) model for describing in-patient flow through a hospital ward. Historical data of 24 wards over the years 2004–2006 were used to analyze the arrival process and length of stay distribution. The assumptions of the model for in-patient flow hold for most clinical wards. Our measure for the goodness of fit of the model, in terms of number of occupied beds, was 0.86 on average (0 = poor, 1 = perfect), indicating that the Erlang loss model is an accurate description of the number of occupied beds. The relation between the size of a hospital unit, the target occupancy rate, and probability of a refused admission is probably the most important lesson learned.

Then this standard queuing model was implemented in MS Excel to make a decision support system, which is both user-friendly and technologically accessible. With this tool one can easily evaluate the current size of nursing units. Also, the effect of merging departments on operational efficiency can be quantified. In a hospital where wards are relatively small and distributed over the different medical disciplines it is practically impossible to operate at both an acceptable service level, in terms of low blocking probability, and a high occupancy rate. When budgets get tight and the number of hospital admissions is increasing we have to seriously investigate the benefits of merging departments or mixing patient flows. Of course, due to the far-reaching specialization, especially in university hospitals, this merging has a limit. Nevertheless, we believe there is still a great potential in efficiency gain for most hospitals.

Appendix A: Statistical test for Poisson arrivals

Here we present some details regarding formal tests to assess the goodness of fit of the Poisson distribution for the number of scheduled and unscheduled arrivals.

Scheduled admissions To formally assess the goodness of fit, we performed the Pearson's chi-square test for hematology 2004 and the 24 wards based on the data of 2005. Details on the chi-square test can be found in many textbooks on statistics (Rice 1995). For hematology 2004 the p-value is 0.318 giving little reason to doubt the Poisson assumption. For 2005, the null hypothesis that the number of arrivals stem from a Poisson distribution is not rejected for 11 out of the 24 wards at a confidence level $((1-\alpha)\cdot 100\%)$ of 95%. For $\alpha = 2.5\%$, the null hypothesis is not rejected for 15 wards.

We note that it is also possible to use a test based on the Poisson index of dispersion (Fisher 1950; Rice 1995), which is closely related to the variance/mean ratio described in Sect. 3. Specifically, under the null hypothesis, the dispersion index $((n-1)\cdot \text{variance}/\text{mean})$ has a chi-squared distribution with $n-1$ degrees of freedom, where n is the number of observations. For 2005, $n = 260$ representing the number of weekdays. Using the Central Limit Theorem and the Normal distribution function, the approximate confidence interval for the variance/mean ratio is (0.828, 1.172) for the two-sided test with $\alpha = 5\%$. In this case, the null hypothesis is not rejected for 12 of the 24 wards. Note that for the one-sided test with $\alpha = 5\%$ the null hypothesis is rejected when the index of dispersion is in (0, 0.855), which is the case for 9 wards.

Unscheduled admissions The Pearson's chi-square goodness of fit test for the medical intensive care unit 2005 gives a p-value of 0.587, giving little reason to doubt the Poisson assumption. For most of the 24 wards the Poisson distribution provides a good fit for the number of unscheduled arrivals. For $\alpha = 5\%$ the null hypothesis of a Poisson distribution is rejected for 2 wards, whereas this happens for 1 ward for $\alpha = 2.5\%$.

References

Bekker, R., & de Bruin, A. M. (2009). Time dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*. doi:10.1007/s10479-009-0570-z.

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

de Bruin, A. M. Nijman, B. C., Caljouw, M. F., Visser, M. C., & Koole, G. M. (2007). Bedden Beter Bezet. *Zorgvisie*, *4*, 29.

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

Fisher, R. A. (1950). The significance of deviations from expectation in a Poisson series. *Biometrics*, *6*, 17–24.

Gallivan, S. (2008). Challenging the role of calibration, validation and sensitivity analysis in relation to models of health care processes. *Health Care Management Science*, *11*, 208–213.

Gallivan, S. et al. (2002). Booked inpatient admissions and hospital capacity: mathematical modelling study. *British Medical Journal*, *324*, 280–282.

Gini, C. (1912). Variabilità e mutabilità contributo allo studio delle distribuzioni e delle relazione statistiche. *Studi Economico-Giuredici dell' Università di Cagliari 3 (Part 2)*, *i–iii*, 3–159.

Green, L. V. (2002). How many hospital beds? *Inquiry—Blue Cross and Blue Shield Association*, *39*, 400–412.

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

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

Kleijnen, J. P. C., Cheng, R. C. H., & Bettonvil, B. (2000). Validation of trace-driven simulation models: more on bootstrap tests. In *Proceeding of the 2000 winter simulation conference* (pp. 882–892).

Koizumi, N., Kuno, E., & Smith, T. E. (2005). Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, *8*, 49–60.

Li, X., Beullens, P., Jones, D., & Tamiz, M. (2008). An integrated queuing and multi-objective bed allocation model with application to a hospital in China. *Journal of Operational Research Society*. doi:10.1057/palgrave.jors.2602565.

Little, J. D. C. (1961). A proof of the queueing formula $L = \lambda W$. *Operations Research*, *9*, 383–387.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*, 209–219.

Rice, A. R. (1995). *Mathematical statistics and data analysis*. N. Scituate: Duxbury Press.

Walley, P., Silvester, K., & Steyn, R. (2006). Managing variation in demand: lessons from the UK National Health Service. *Journal of Healthcare Management*, *51*, 309–322.

Whitt, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal*, *63*, 689–708.

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

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