Journal of the Operational Research Society (2002) 53, 19–24
www.palgrave-journals.com/jors

A queueing model for bed-occupancy management and planning of hospitals

F Gorunescu[1], SI McClean[2]* and PH Millard[3]

[1]*University of Medicine and Pharmacy, Craiova, Romania;* [2]*University of Ulster, Northern Ireland, UK; and* [3]*St George's Hospital Medical School, London, UK*

The aim of this paper is, on the one hand, to describe the movement of patients through a hospital department by using classical queueing theory and, on the other hand, to present a way of optimising the use of hospital resources in order to improve hospital care. A queueing model is used to determine the main characteristics of the access of patients to hospital, such as mean bed occupancy and the probability that a demand for hospital care is lost because all beds are occupied. Moreover, we present a technique for optimising the number of beds in order to maintain an acceptable delay probability at a sufficiently low level and, finally, a way of optimising the average cost per day by balancing costs of empty beds against costs of delayed patients.
Journal of the Operational Research Society (2002) **53,** 19–24. DOI: 10.1057/palgrave/jors/2601244

Keywords: queueing theory; geriatric medicine; bed occupancy; hospital costs model

Introduction

An under-provision of hospital beds leads to patients in need of hospital care being turned away, and the build-up of waiting lists or stress on another part of the hospital system. For example, when insufficient medical beds are provided to meet demand, emergency medical patients spill over into surgical beds; consequently, surgical waiting lists increase as planned admissions are postponed. On the other hand, an over-provision of hospital beds is wasteful of scarce resources. The aim of this paper is to show how a classical queueing model may be used to optimise the allocation and use of hospital beds in order to improve patient care.

Queueing models are widely used in industry to improve customer service, for example in supermarkets, banks and motorway toll booths. In a market economy, customers who have poor service go elsewhere; however, when hospital beds are unavailable, patients have no option but to wait at home, in accident and emergency departments or wards which are inappropriately staffed, possibly without access to appropriate specialised equipment. Previous work has also applied queueing theory to health service applications.[1–5]

Compartmental models have previously been shown to provide a valid description of patient movements through the hospital system. Such models may be used to plan health and social care for an ageing population where 'compartments' may correspond to short (acute), medium (rehabilitative) or long stay care. A two-compartment discrete-time deterministic model to describe such movements of patients has been developed by Harrison and Millard[6] and shown to explain the empirical result of Millard[7] that the distribution of length of stay of patients in a geriatric department is described by a mixed exponential distribution. An extension of this model to its continuous-time stochastic analogue is to be found in Irvine, McClean and Millard.[8] Using the methodology of Harrison and Millard,[6] Taylor *et al*[9] have developed a four-compartment discrete-time deterministic model of patient behaviour; in Taylor, McClean and Millard,[10] five and six compartments were also considered.[11]

Such compartmental models may all be regarded as *phase-type distributions*[12,13] which describe the time to absorption of a finite Markov chain in continuous time, when there is a single absorbing state and the stochastic process starts in a transient state. Distributions of this form have considerable generality, and include exponential (single phase), Erlang, and mixed exponential distributions; indeed, any continuous distribution with non-negative support can be arbitrarily closely approximated by one of phase-type form. This allows us to capture heterogeneity in patient lengths of stay where there may be a large variation in the amount of time patients spend in hospital and, as is the case for geriatric medicine, some patients may stay in hospital for very long periods.

As regards the admission policy in the geriatric department, previous work has considered the situation where admissions occur at random,[8,10] ie Poisson arrivals, and shown that such an assumption is reasonable for a stable hospital system. Such an admission process can be envisaged as a superposition of random events, which happen rarely to a large population, eg patients from the pool of

**Correspondence:* S McClean, School of Information and Software Engineering, Faculty of Informatics, University of Ulster, Cromore Road, Coleraine BT52 ISA, Northern Ireland, UK.
E-mail: si.mcclean@ulst.ac.uk

people in the catchment area fall ill from time to time. Such a scenario is well known to result in a Poisson process.[14]

This paper uses results from queueing theory to determine optimal bed numbers for a hospital system where we describe patient arrivals by a Poisson process, hospital beds are the servers and lengths of stay are modelled using phase-type distributions. In queueing terminology this is known as a M/PH/c queue[15,16] where M denotes Poisson (Markov) arrivals, the service distribution is phase-type, and c is the number of servers (the beds). It is also assumed that the queueing system is in steady state which, in practical terms, means that we assume that the hospital system has been running, in its present form, for a few years. The aim is to provide decision-support methods, including costing, for planning services within a hospital.[17,18]

The theoretical model

We consider a M/PH/c queue in which the number of beds is fixed and no queueing is allowed; a patient finding upon arrival that all c beds are occupied, is lost; in reality these patients are often admitted to the beds of other specialities, or they may wait at home or in the Accident and Emergency department until beds become available. The general problem is therefore rather complex; for the moment we focus on the simpler model, which does not consider such alternatives, in the interests of providing clarification of the issues involved. We assume that patient arrivals follow a Poisson process with rate λ and the service time is Phase-type with pdf

$$f(t) = \sum_{i=1}^{k} \alpha_i \rho_i e^{-\alpha_i t} \quad \text{where} \quad \sum_{i=1}^{k} \rho_i = 1 \tag{1}$$

and corresponding mean:

$$\tau = \sum_{i=1}^{k} \frac{\rho_i}{\alpha_i} \tag{2}$$

Here the various parameters, k (the number of phases/compartments), the α_is (mixing proportions), and the ρ_is (transition rates), may be estimated using likelihood ratio tests.[13,19] The average number of arrivals occurring during an interval of length t is λt and, therefore, the average number a of arrivals during an average length of stay τ is $a = \lambda\tau$, known as the *offered load*.

Applying standard results from queueing theory,[15,16] we see that the probability of having j occupied beds is given by

$$p_j = \frac{a^j/j!}{\sum_{k=0}^{c} a^k/k!} \tag{3}$$

a well-known result which asserts that the statistical equilibrium probability p_j depends on the service-time distribution only through its mean. Let us recall that the system is said to be in *statistical equilibrium* if, after a sufficiently long period of time, the state probabilities are independent of the initial conditions. From the above formula we deduce that the probability that all c beds are occupied or the fraction of arrivals that is lost, is given by *Erlang's loss formula*:

$$B(c, a) = \frac{a^c/c!}{\sum_{k=0}^{c}(a^k/k!)} \tag{4}$$

An approximation to the desired value of $B(c, a)$ may be computed by using the normal approximation to the Poisson distribution where we denote the standard normal probability function by $\Phi(x)$. Then:

$$\sum_{j=c}^{\infty} \frac{(\lambda\tau)^j}{j!} e^{-\lambda\tau} \approx 1 - \Phi\left(\frac{c - \lambda\tau - 0.5}{\sqrt{\lambda\tau}}\right)$$

provided the offered load $\lambda\tau$ is not too small relative to the capacity c.

Another useful quantity is represented by the mean number of occupied beds:

$$a' = a[1 - B(c, a)] \tag{5}$$

also known as the *carried load*. It is easy to see that the offered load a is the load that would be carried if the number of beds were infinite and the carried load a' is just that portion of the offered load that is not cleared (lost) from the system.

The average time spent in the geriatric department by an arbitrary patient is then the average length of stay multiplied by the probability of being admitted, namely:

$$W = \tau[1 - B(c, a)] \tag{6}$$

In order to measure the degree of utilisation of a group of beds, we define the bed occupancy by

$$\rho = \frac{a'}{c} \tag{7}$$

clearly, we must always have $\rho \leqslant 1$, otherwise the system cannot be in steady state.

Optimising the number of beds

The hospital system that we have described can result in a patient being turned away because all beds are occupied; such a patient may not receive the necessary care. On the other hand, the goal of the hospital is to assign beds in order to provide the best level of service possible. We here address this dilemma by minimising the number of beds subject to maintaining the delay probability at a (specified) sufficiently low level. Our goal is, therefore to provide an inverse function which gives the value of c corresponding to a given delay probability $B(c, a)$.

With this aim in view, let us consider a M/PH/c queue, with offered load $a = \lambda\tau$. Suppose we wish to determine the minimum number of beds c in order to achieve a delay probability of not more than a pre-specified value

$B(c, a) = v$. We then wish to determine c_0 which is the smallest value of c for which:

$$B(c, a) = \frac{a^c/c!}{\sum_{k=0}^{c}(a^k/k!)} \geqslant 1 - v \tag{8}$$

where v is chosen as the highest proportion of refused patients which we are prepared to tolerate. We are here calculating the inverse of the formula in the previous section. Thus we specify the maximum acceptable delay probability and calculate the corresponding number of beds; previously, we specified the number of beds and calculated the corresponding delay probability.

Optimising the average cost per unit time

The main goal of a geriatric department is, on the one hand, to assign the best service to its patients, seen in this context as the minimum delay, and, on the other hand, to maintain a maximum utilisation of resources, in this case the beds. To determine the optimal policy we use a notion from inventory theory known as the base-stock policy.[6] Such optimal inventory models commonly are concerned with the trade-off between holding costs of (expensive) inventory items and penalty costs if a demand is left unsatisfied.[20] The so-called *base-stock policy* is often used in inventory systems of expensive and slow-moving items for which unit demands occur; under this control rule, a replenishment order for exactly one unit is placed each time the on-hand inventory decreases by one unit on the occurrence of a demand. Demands occur according to a Poisson process, rate λ. The base-stock level c represents the optimum inventory level that balances inventory costs against unsatisfied demands if we run out of stock. Applying this approach to our situation, let us assume that the total number of beds in the geriatric compartment is c, where c equals the number of occupied beds plus the number of idle beds.

Here, beds correspond to the inventory where idle beds are the on-hand inventory and occupied beds are unfilled orders. A patient arrival corresponds to a demand and the subsequent length of stay of the patient corresponds to waiting time for replenishment. Patients who are turned away because there are no empty beds correspond to unsatisfied demands. In a similar fashion to the inventory problem we envisage a *holding cost* of $h > 0$ per day for each empty bed, while a fixed *penalty cost* of $\pi > 0$ is incurred for each patient that is turned away (lost demand); in general, we also envisage a profit $p > 0$ per patient per day. The total cost to the customer (health service purchaser, insurance company or the patient himself) is therefore the sum of variable costs (treatment), fixed costs (holding costs), profit, and penalty costs (of turning away patients). However, we envisage a situation where the variable costs are paid directly by the customer while the fixed costs and penalty costs are incurred by the service provider; in what follows we therefore ignore variable costs which are not incurred by the service provider—we then term the total cost minus the fixed costs, the actual cost.

We are interested in how to choose the total number of beds c in the geriatric department as a function of the parameters λ, τ, h, π and p in order to minimise the long-run average cost per unit time. To model this situation we consider a M/PH/c queueing system where λ and τ are as defined previously and $a = \lambda\tau$. As was observed earlier, the average demand that is lost per unit time equals $\lambda B(c, a)$, the average idle-bed inventory equals $c - a[1 - B(c, a)]$ (using Equation (5)) and, therefore, we can conclude that the average service provider revenue per day under the base-stock policy with bed level c is given by:

$$r(c) = -\pi\lambda B(c, a) - h\{c - a[1 - B(c, a)]\} + pa[1 - B(c, a)] \tag{9}$$

We may therefore find c to maximise $r(c)$; this corresponds to the optimal number of beds where we balance the number of empty beds against the number of delayed patients. We note that, in Equation (9), for a public health service where $p = 0$, it is the ratio π/h which determines the optimum; here the absolute value of h is just a scaling factor for the average cost per day and we are interested in maximising revenue, or alternatively minimising cost, which is given by

$$g(c) = \pi\lambda B(c, a) + h\{c - a[1 - B(c, a)]\} \tag{10}$$

In what follows we will focus on optimising costs for such a publicly funded health service, where cost minimisation entails finding the optimal balance between holding costs (keeping empty beds available to meet demands) and penalty costs (avoiding having to turn away patients).

Thus far we have considered the problem of finding the number of beds c that minimises the cost for a given arrival rate and average length of stay. Having carried out this optimisation we now want to evaluate the sensitivity of this result to changes in the arrival rate and/or average length of stay. In general, the average cost per day $g(c)$ may not be very sensitive to the exact optimal value of c and it may be difficult to determine exact values for the holding costs h penalty costs π and profit p in a non-commercial environment. In such situations the so-called *indifference curves*[21] can be used to show under what conditions we are indifferent between neighbouring values of the control variable c. Therefore, we say we are indifferent between the use of the consecutive c and $c + 1$ when

$$g(c) = g(c + 1).$$

By simple calculation we deduce that this equation is equivalent to

$$\lambda\tau\{B(c, \lambda\tau) - B(c + 1, \lambda\tau)\} = \left(1 + \frac{\pi}{h\tau}\right)^{-1} \tag{11}$$

This equation shows that the optimal choice of c depends only upon the parameters $\lambda\tau$ and $\pi/(h\tau)$. Consequently, for a

given value of c we may evaluate the indifference by means of a two-dimensional graph of $\pi/(h\tau)$ versus $\lambda\tau$; this graph, which describes Equation (9), is known as the indifference curve.[14] When the curves for different values of c are close together we say that we are indifferent to the choice of π/h for given values of λ and τ. This means that, for such choices of c, the optimal number of beds is not very dependent on the ratio π/h. This is a useful observation since in practice the precise value of the penalty cost π may be difficult to specify.

Application of the model

The loss model

We illustrate the methodology using bed occupancy data collected at the Department of Geriatric Medicine of St. George's Hospital, London in January 2000.[22] This Department provides an acute, rehabilitative and long-stay service. The data, which included 140 patients with an average age of 84 years, were found to be well described by a three-phase model. We consider the model for these data where the arrival rate was $\lambda = 5.9$ patients per day and the mean length of stay $\tau = 24.9$ days. In Table 1 we present a table of various system characteristics as a function of the number of beds c for this department. Thus, when there are 120 beds, 21% of patients are turned away from the geriatric department, and on average there are 116 patients occupying beds, representing 97% bed occupancy. We can see from this table that, as we would expect, the greater the number of beds, the lower the probability of a patient being turned away (lost demands) and the lower the bed occupancy. A graph of percentage of patients turned away against number of beds is provided in Figure 1.

Table 1 The main service features ($\lambda = 5.9$ patients per day, $\tau = 24.9$ days, $\lambda\tau = 146.91$ patients)

Number of beds c	*Lost demands probability* $B(c, \lambda\tau)$(%)	*Mean number of patients (L)*	*Bed occupancy* ρ (%)
120	21	116	97
125	18	120	96
130	15	125	96
135	12	129	96
140	9	134	95
145	7	137	94
150	5	140	93
155	3	143	92
160	2	144	90
165	1	145	88
170	1	145	86
175	0	147	84

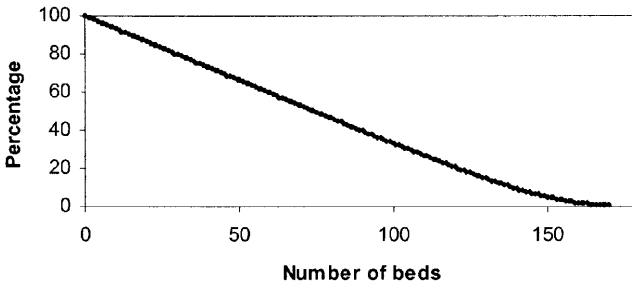

Figure 1 Percentage of patients turned away.

The optimisation method

As an illustration of the optimisation method for the number of beds, we consider some typical values for the delay probability $B(c, a)$. In Table 2 we present suitable values of B along with the corresponding number of beds needed to maintain this level of service. Thus, for example, if we wish to ensure that at most 5% of patients are turned away, we must have at least 150 beds in the geriatric department.

The cost model

In what follows we assume that the profit is zero since the data are from the National Health Service. In order to illustrate the optimal base-stock level, ie the average cost per unit time for holding and penalty costs, as a function of the number of beds c, we have presented in Table 3 the values of $g(c)$ corresponding to different ratios of the penalty costs π to the holding cost h. Here we assume that the total cost per patient per day is £168 where £50 are incurred with respect to the bed and £118 with respect to the treatment; since the data are from the National Health Service we assume that the profit is zero. We then estimate the holding cost as $h =$ £50 per day and the penalty cost as 25% of the total cost of turning away the patient, in this case taken as the cost per day multiplied by the expected length of stay, ie $\pi = 168 \times 24.9 \times 0.25 =$ £1046. This approach is meant to be indicative of a ballpark figure for cost and is based on an assumption that penalty may be regarded in some sense as lost revenue incurred when a patient is turned away due to there being no empty beds available; total lost revenue per patient turned away is then cost per day multiplied by expected number of days that have been lost. This figure is then adjusted to take account of the fact that a proportion of revenue must balance costs, so it is really the profit that is affected by lost patients. We are here calculat-

Table 2 Number of beds and queue characteristics corresponding to $B(c, \lambda\tau)$

Delay probability $B(c, \lambda\tau)$	0.1%	1%	5%	10%
Minimum number of beds	179	166	150	139

Table 3 The values of the average cost per unit time £$g(c)$

Beds c	$\pi/h = 10$ $g_1(c)$	$\pi/h = 20$ $g_2(c)$	$\pi/h = 30$ $g_3(c)$	$\pi/h = 40$ $g_4(c)$
120	781	1390	1999	2608
125	723	1244	1765	2286
130	676	1112	1548	1984
135	643	998	1353	1708
140	629*	908	1187	1466
145	638	848	1058	1268
150	677	827*	976	1126
155	752	851	951*	1050
160	867	927	988	1049*
165	1022	1055	1089	1122
170	1212	1229	1245	1262

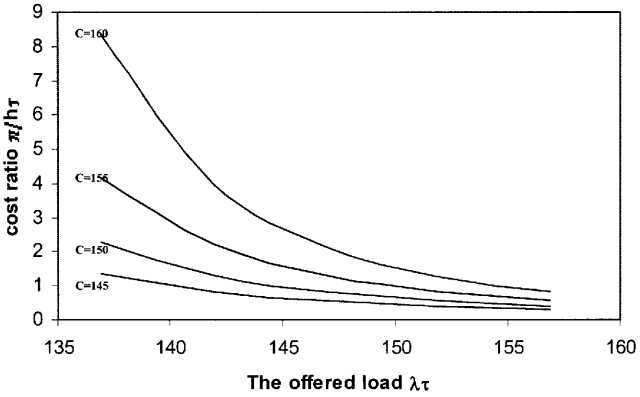

Figure 3 The indifference curves.

ing penalty as the notional profit that would typically have been produced if the hospital were private; in practice, the penalty costs for a publicly funded service are additional costs to the care system and include costs of patient distress and premature dependency or death.

In Figure 2 we have drawn the corresponding graph of $g(c)$. Here the offered load is $a = 146.9$ and we have marked by * in Table 3 the minimum value of the average cost per unit time. We note from Table 3 that if we increase the penalty to holding cost ratio π/h four times, from 10 to 40, then the corresponding feedback of the number of beds needed to obtain minimal costs indicates an increase of only 14% from 140 to 160, suggesting that the ratio π/h, has no significant influence on the optimal number of beds.

Finally, in Figure 3 we have drawn the indifference curves for different inventory levels $c = 145$, 150, 155 and 160. This picture suggests that we may be indifferent to the ratio π/h if the number of beds is 145, 150 and even 155 (the lower curves) but, when the number of beds exceeds 160, then the cost changes dramatically. This is a reflection of the rapidly increasing costs for more than 160 beds in Figure 2.

Conclusions

The methodology we have developed enables estimation of the main characteristics of access to service for patients and hospital managers: the probability of lost demands, and the mean number of patients in the hospital department (bed occupancy). In addition, the model enables the hospital manager to balance the cost of empty beds against the cost of turning patients away, thus facilitating a good choice of bed provision in order to have low cost and high access to service. We have thus provided a means of calculating optimal bed numbers given an acceptable level of patient rejection. Our approach has been illustrated using data for a geriatric hospital department. The results agree with anecdotal findings[23] that running with percentage occupancy above the high eighties leads to significant increases in rejection of patients.

Figure 2 Actual cost per day.

We have assumed that the hospital department may be described by a M/PH/c queue in steady state. This enabled us to build on previous work that established that a Phase-type distribution gives a good fit to the length of stay of such patients and enables us to estimate the arrival rate λ and average length of stay τ using a midnight bed census. However, more generally, the queueing theory results used in this paper are valid for any length of stay distribution so all we need to use the results is an estimate for the arrival rate and the average length of stay. The advantage of using the phase-type model for length of stay distribution is that it provides a useful description of the data and a convenient way of estimating arrival rates; however the phase-type assumption is not essential to the approach. Thus, provided that the Poisson arrival and steady state distribution hold, our methodology is valid. Previous work[24] has suggested that, while there is some seasonality in patient admissions, the Poisson assumption is appropriate for most of the time. Steady state is likely to be appropriate for a department that has been stable for a few years. For example, a simulation study[25] illustrated that, for a hospital department starting with no patients, steady state in most cases took a few years to achieve; patients with very long lengths of stay took up to 5 years to achieve steady state.

We have here taken the penalty costs as would-be lost profit if the patient is turned away from the department. More generally, penalty costs may include costs of unused

resources, eg staff costs, costs to other parts of the health service—a patient turned away by one department may have to be cared for by another less suitable one, costs to social services who may have to maintain the patient at home, and costs of human suffering. However, such costs may be hard to specify while our current approach to costing is easy to quantify. In addition our sensitivity analysis has suggested that the optimal cost may not be very sensitive to the ratio of penalty cost to holding cost.

It must be emphasised that our results are intended to be indicative of general patterns rather than descriptive of exact behaviour, in the interests of providing clarification of the issues involved. Thus while there are times when seasonal effects, such as an influenza epidemic, may mean that the Poisson assumption no longer holds, it is likely that our assumptions broadly hold for a hospital department in normal operation, most of the time. In this case, all that is required to use these results is an estimate of patient arrival rate, including patients who are not admitted, and average length of stay in the department. We may then use our methodology to estimate the average bed occupancy and determine an optimal number of beds for the department. As such our approach is likely to prove an extremely useful tool for hospital managers.

References

1 Worthington DJ (1987). Queueing models for hospital waiting lists. *J Opl Res Soc* **38**: 413–422.
2 Worthington DJ (1991). Hospital waiting list management models. *J Opl Res Soc* **42**: 833–843.
3 Worthington D and Wall A (1999). Using the discrete time modelling approach to evaluate the time-dependent behaviour of queueing systems. *J Opl Res Soc* **50**: 777–788.
4 Liu L and Liu X (1998). Block appointment systems for outpatient clinics with multiple doctors. *J Opl Res Soc* **48**: 1254–1259.
5 Lehaney B, Clarke SA and Paul RJ (1999). A case of an intervention in an outpatients department. *J Opl Res Soc* **50**: 877–891.
6 Harrison GW and Millard PH (1991). Balancing acute and long-stay care: the mathematics of throughput in departments of geriatric medicine. *Meth Inform Med* **30**: 221–228.
7 Millard PH (1988). Geriatric medicine: A new method of measuring bed usage and a theory for planning. MD thesis, University of London.
8 Irvine V, McClean S and Millard P (1994). Stochastic models for geriatric in-patient behaviour. *IMA J Math Appl Med Biol* **11**: 207–216.
9 Taylor G, McClean S, and Millard P (1996). Geriatric patient flow-rate modelling. *IMA J Math Appl Med Biol* **13**: 297–307.
10 Taylor G, McClean S and Millard P (1998). Continuous-time Markov models for geriatric patient behaviour. *Appl Stoch Models Data Anal* **13**: 315–323.
11 Taylor G, McClean S and Millard P (2000). Stochastic models of geriatric patient bed occupancy behaviour. *JRSS Ser A* **163**: 39–48.
12 Neuts MF (1981). *Matrix Geometric Solution in Stochastic Models*, Johns Hopkins University Press: Baltimore, MD.
13 Faddy M (1994). Examples of fitting structured phase-type distributions. *Appl Stoch Models Data Anal* **10**: 247–255.
14 Cox DR (1962). *Renewal Theory.* Methuen: London.
15 Cooper RB (1972). *Introduction to Queueing Theory.* McMillan: New York.
16 Tijms HC (1986). *Stochastic Modelling and Analysis. A Computational Approach.* Wiley: Chichester.
17 Beech R, Brough RL and Fitzsimons BA (1990). The development of a decision-support system for planning services within hospitals. *J Opl Res Soc* **41**: 995–1006.
18 McClean S, McAlea B and Millard P (1998). Using a Markov reward model to estimate spend-down costs for a geriatric department. *J Opl Res Soc* **10**: 1021–1025.
19 Faddy MJ and SI McClean (1999). Analysing data on lengths of stay of hospital patients using phase-type distributions. *Appl Stoch Models Data Anal* **15**(4): 311–317.
20 Stevenson WJ (1996). *Production/Operations Management*, 6th edn. Irwin, McGraw-Hill, USA.
21 Silver EA and Smith SA (1977). A graphical aid for determining optimal inventories in a unit inventory replenishment system. *Mngt Sci* **24**: 358–359.
22 Vinicchayakul, R (2000). Costing care in geriatric medicine. MSc dissertation, University of London.
23 Horrocks P (1986). The components of a comprehensive district health service for elderly people: a personal view. *Age Ageing* **15**: 321–342.
24 McClean SI and Millard PH (1993). Modelling in-patient bed usage behaviour in a department of geriatric medicine. *Meth Inform Med* **32**: 79–81.
25 El-Darzi E, Vasilakis C, Chaussalet T and Millard P (1998). A simulation modelling approach to evaluating length of stay, occupation, emptiness and bed-blocking in a hospital geriatric department. *Health Care Mngt Sci* **2**: 143–149.

Received December 1998;
accepted June 2001 after three revisions