

Modelling Activities at a Neurological Rehabilitation Unit

J D Griffiths, J E Williams, R M Wood*

School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, United Kingdom, CF24 4AG

** Corresponding author: Dr Richard Wood, woodrm@cardiff.ac.uk*

Abstract

A queuing model of a specialist neurological rehabilitation unit is studied. The application is to the Neurological Rehabilitation Centre at Rookwood Hospital (Cardiff, UK), the national rehabilitation unit for Wales. Due to high demand this 21 bed inpatient facility is nearly always at maximum occupancy, and, with a significant bed-cost per day this makes it a prime candidate for mathematical modelling. Central to this study is the concept that treatment intensity has an effect on patient length of stay. The model is constructed in four stages. First, appropriate patient groups are determined based on a number of patient-related attributes. Second, a purpose-built scheduling program is used to deduce typical levels of treatment to patients of each group. These are then used to estimate the mean length of stay for each patient group. Finally, the queuing model is constructed. This consists of a number of disconnected homogeneous server queuing systems; one for each patient group. A Coxian phase-type distribution is fitted to the length of time from admission until discharge readiness and an exponential distribution models the remainder of time until discharge. Some hypothetical scenarios suggested by senior management are then considered and compared on the grounds of a number of performance measures and cost implications.

Keywords: *Queuing theory, Markov modelling, Scheduling, OR in health services*

1. Introduction

Neurological rehabilitation¹ is the process that aids functional recovery following an injury to the brain. Its primary goal is to ‘*enable the patient to function at the highest level of independence possible, given his or her disability*’ (Hanlon, 1996). A patient typically begins a programme of rehabilitation after a period of acute care following the initial onset of the injury. Acquired Brain Injury (ABI) is defined in Turner et al, 2002 as ‘*an injury to the brain that has occurred since birth*’ and can result from traumatic (e.g. road accident) or non-traumatic (e.g. tumour) causes. The severity of ABI can be classified as mild, moderate or severe (Teasell et al, 2007). In the UK it is estimated that moderate to severe ABI occurs in

¹ See Barnes, 2003 for a review of basic principles.

25 per 100,000 people per year (Turner-Stokes, 2003). It is such patients that are appropriate candidates for an inpatient rehabilitation programme.

The aim of this study is to mathematically model the activities at an inpatient neurological rehabilitation unit. The application of this project is the Neurological Rehabilitation Centre at Rookwood Hospital in Cardiff, Wales. Admission to this facility is by referral only. There are approximately five referrals received every fortnight and these are placed in a non-pre-emptive priority²-based queue. This level of demand exceeds supply and so there is always a queue. The size of the queue is reduced by either the withdrawal of a referral (due to transfer, discharge or death) or by an admission. For a patient to be admitted there must be an available bed and a sufficient skill mix on-hand to cater for the needs of the referred patient (it is therefore not always possible to admit the highest priority patient in the queue). Because of this, there are typically only 21 beds occupied out of an available 25. Once admitted, treatment is provided by (up to) six departments³ in line with the weekly master schedule. When patients have demonstrated satisfactory recovery in the field of each department, a date of expected discharge readiness is set and the discharge destination⁴ is informed of this. The date of readiness is generally about four months from admission. However, this is rarely the date of ultimate discharge due to the inability of the discharge destination to accept the patient on-time. This delay to discharge or bed-blocking typically adds about one month to length of stay.

What has been explained in the previous paragraph is a queuing system. This can be modelled by discrete-event simulation (to capture the precise time and state relationships of events) or queuing theory. The latter is used in this study owing to the research potential and the benefits of analytic over simulated results. Despite a thorough literature search, no studies were found that produce a queuing model of a neurological rehabilitation facility. This lack of previous research background alongside the cost (£480 bed cost per day) and scarcity (annual demand = 130; annual supply = $21 \times 12/5 = 50$) of services, provide a strong motivation for this study.

Whilst there is no specific literature linking queuing theory and neurological rehabilitation, there are some queuing studies in geriatric care that are comparable. The congruency in

² Determined by injury severity and time spent in queue.

³ Physiotherapy, occupational therapy, speech and language therapy, psychology, dietetics, orthopaedics (in order of decreasing size as measured by number of full-time equivalents).

⁴ Typically (62%) own home with social services support (see Table A.1).

patient flow permits a comparison, i.e. accident and emergency followed by some form of inpatient facility and then community care. Taylor et al, 1998 use a queuing model to study the movement of geriatric patients between the hospital and the community. Faddy & McClean, 2005 fit phase-type distributions to length of stay in the hospital and community and later, in McClean & Millard, 2007, incorporate this service distribution within a queuing model. The authors state that *'keeping patients longer in hospital, and improving their fitness for discharge can reduce the transfer of patients to secondary care systems and may in the long-term both improve hospital performance and reduce costs'*. This very notion is echoed in Turner-Stokes, 2007 with regard to the rehabilitation of ABI patients.

However, it is not just length of stay (LOS) that affects functional ability on discharge. Many clinical studies indicate that the intensity of treatment received also has an effect on patient outcome⁵ (Heinemann et al, 1995, Cifu et al, 2003). There are also those that suggest that treatment intensity affects LOS (Blackerby, 1990, Slade et al, 2002) and since LOS affects outcome (Spivack et al, 1992, Sandhaug et al, 2010) a trilateral relationship between these three measures is inferred. Unfortunately, for this study, data is not available for community care and so the queuing model is restricted to the rehabilitation unit only. There is therefore no reason to permit variability on outcome within this study and so treatment intensity and LOS are the respective independent and dependent variables. In the queuing model, average LOS is equivalent to the mean service time of a service channel (bed). The mean service rate is the reciprocal of this value. The effect of treatment intensity is incorporated by conditioning these service rates on the intensity of treatment received.

However, the amount of treatment provided is not directly controlled by clinicians – it depends on the amount allocated on the weekly master schedule⁶. In order to ask realistic questions of the model that relate to changes in treatment intensity, it is necessary to produce appropriately modified versions of this master schedule. For example, if the workforce is increased by 50% then a master schedule based on this change would be constructed from which treatment intensity can be deduced. Since the construction of the master schedule by hand takes many hours each week it is not appropriate to ask therapists to produce these versions. An automated scheduling program that replicates the process is therefore sought. However, such a program is not suitable for all of the departments. In fact, only the

⁵ Defined as functional ability on discharge.

⁶ Departments schedule their treatment consecutively and in order of increasing size during the preceding week. The master schedule is an amalgamation of these departmental timetables.

physiotherapy department has a sufficient number of staff and a clear-enough set of treatment allocation guidelines to make an automated program viable. It is assumed that any changes made to the physiotherapy department are replicated across the other departments. The quality of timetables produced by the program actually surpassed those produced by hand and so replaced the manual approach for the week-to-week scheduling of physiotherapy treatment. This program is described in Section 2.

The rest of this paper is structured as follows. A computer program designed to solve a multi-server queuing system with balking and renegeing for steady-state results is described in Section 3. This will be of use in the construction of the model (Section 5) in which a Coxian distribution is fitted to the length of time from admission to discharge readiness and an exponential distribution is fitted to the remainder of time until discharge. Thereafter, a number of hypothetical scenarios are considered using the model in Section 6. A discussion is provided in the final section.

2. The automated scheduling program

2.1 Background

Treatment is provided to patients in pre-arranged sessions. For the physiotherapy department there are about ten types of session that can be performed. These differ in duration, activity performed, and number of therapists/patients in attendance. For example, a ‘stretch’ is 30 minutes with one or two therapists and a single patient, whilst a ‘group’ is 45 or 60 minutes with one therapist and about three patients. A ‘standard’ session is 45 or 60 minutes and is either a ‘single’, ‘double’ or ‘triple’ depending on how many therapists are involved in treating the patient.

The physiotherapy timetable details the start and end time, type, providing therapist(s), and recipient(s) of each session scheduled for that week. It is produced on the Thursday of the preceding week (to that for which treatment is being scheduled) after all of the other (smaller) departments have allocated their sessions. Treatment allocation is dependent on the following six variables: patient demand (number of requests for each session type), patient attributes (preferred therapist, preferred time of day, number of therapists to stretch etc), patient priority (scale of zero to five), patient availability (after sessions allocated from other departments), therapist attributes (band level, ability to perform certain session types) and therapist availability (based on working hours, annual leave, meetings etc).

2.2 Constraints

A number of constraints have been deduced through discussion with clinicians.

2.2.1 Hard constraints

There are 17 hard constraints. These range from rather obvious ones like *'the start time of a session must be before the end time'* to *'the band level of at least one providing therapist must be at least the patient's minimum requirement'*.

2.2.2 Moderate constraints

These can be thought of as hard constraints whose satisfaction is non-essential. It is thought that by representing the most important soft constraint violations by a discrete measure a more precise picture of their incidence can be formed. This will allow for a more targeted approach in their satisfaction. There are four moderate constraints. One of these relates to the importance of ensuring an even spread of treatment sessions throughout the working week. For group sessions, the number of violations is determined as follows. First, the minimum, $n^{\min} = \left\lfloor \frac{1}{5} \sum_{\forall d} n_d \right\rfloor$, and maximum, $n^{\max} = \left\lceil \frac{1}{5} \sum_{\forall d} n_d \right\rceil$, number of group sessions allowed to occur on each day are deduced (where n_d is the number of group sessions that occur on day $d = 1, \dots, 5$). Second, each day is scored depending on whether the number of sessions is contained within the acceptable range, i.e.

$$\delta_d^{\min} = \begin{cases} 0 & \text{if } n_d \geq n^{\min} \\ n^{\min} - n_d & \text{if } n_d < n^{\min} \end{cases}, \quad \delta_d^{\max} = \begin{cases} 0 & \text{if } n_d \leq n^{\max} \\ n_d - n^{\max} & \text{if } n_d > n^{\max} \end{cases} \quad (1)$$

Finally, the number of violations is calculated as

$$\max \left(\sum_{\forall d} \delta_d^{\min}, \sum_{\forall d} \delta_d^{\max} \right) \quad (2)$$

There are no constraint violations if the difference in number of group sessions assigned on any two days is not more than one.

2.2.3 Soft constraints

There are five soft constraints whose weighted addition forms the continuous-valued objective function. These measure the extent to which desirable conditions (such as allocation to preferred therapist or time of day) are satisfied. Perhaps the most important is attaining a fair allocation of sessions. The component of the objective function that relates to this is calculated as

$$\sum_{i=1}^n p_i \sum_{s=1}^{10} w_s (n_{i,s}^{\text{dem}} - n_{i,s}^{\text{sch}})^2 \quad (3)$$

where p_i is the priority level of patient $i = 1, \dots, n$, w_s is the weight applied to session type s , and $n_{i,s}^{\text{dem}}$ and $n_{i,s}^{\text{sch}}$ are the number of sessions of type s demanded and scheduled for patient i .

2.3 Solution

A three stage local search based approach is used to provide an approximate solution to what is a hierarchical multi-objective combinatorial optimisation problem.

2.3.1 Stage One: Initial construction

There is a specialised constructive heuristic algorithm for each type of session that produces an initial assignment. These stochastic algorithms create an initial solution that is not just valid but is also high quality. This reduces the workload of subsequent local search algorithms and has been shown to enhance the quality of the final solution (since local search algorithms can struggle in highly constrained situations).

2.3.2 Stage Two: Moderate constraint optimisation

The objective here is to reduce the number of moderate constraint violations whilst respecting hierarchical precedence. Although not referred to at the time, the formulae for the number of violations (Section 2.2.2) have been carefully developed to represent a minimal distance to feasibility. The number of violations represents the minimal number of moves or swaps required to satisfy that moderate constraint for the patient in question. Whilst each moderate constraint has a tailor-made algorithm, the approach is essentially the same. First, a patient is selected that has one or more violations of the moderate constraint in question. Second, the algorithm identifies any ‘problem’ sessions that are causing the violation(s). For each of these, a number of corresponding ‘swap’ sessions are intelligently identified that, when

exchanged with the problem session, result in fewer violations of the moderate constraint. This effectively trims the neighbourhood to consider only improved solutions. Each of these is then evaluated and the best one (based on the hierarchy of constraints) is chosen (i.e. guided steepest descent). This represents one iteration. If no improvement can be found then a lower-priority patient is selected. After all patients have been considered then a lower-order moderate constraint is investigated.

2.3.3 Stage Three: Soft constraint optimisation

A generic local search based algorithm is used in repetition to further optimise the solution with respect to hierarchical precedence. First, a list of all sessions is created. Second, a session is picked at random and exchanged, in turn, with the other sessions. Simulated annealing is then used to determine whether or not to accept a neighbouring solution. If reductions in moderate constraints are possible, then such swaps are accepted regardless of objective function value. If a swap is accepted, then the list of sessions is reset. Otherwise, the session is removed from the list to avoid it being chosen again (i.e. tabu search).

2.4 Results

The automated scheduling program (coded in MS VBA owing to user familiarity) was trialled in the latter quarter of 2010. During this period it was used alongside the manual approach and results were compared (Table 1). Not only was timetable quality considerably better, but employee time expended was significantly lower. By hand, the timetable would take a therapist eight hours each week to produce, but with the automated program, this is reduced to fewer than two. In addition, suitable data for audit purposes and future outcomes research are automatically output. Owing to these significant advantages, the program was adopted by the physiotherapy department in January 2011 and has been used each week since.

Table 1 Comparison between manual and automated (12 hours run time) scheduling on a standard desktop computer

| Week start (2010): | 20 Sep | | 4 Oct | | 1 Nov | |
|------------------------------------------------------------|---------|---------|---------|---------|---------|---------|
| Method: | by hand | program | by hand | program | by hand | program |
| # Hard constraint violations | 0 | 0 | 0 | 0 | 0 | 0 |
| # Moderate constraint violations | 12 | 1 | 5 | 0 | 4 | 0 |
| Objective function | 280,435 | 147,385 | 207,720 | 125,350 | 318,050 | 151,765 |
| Avg # clinical minutes demanded for each patient | 279 | 279 | 290 | 290 | 258 | 258 |
| Avg # clinical minutes scheduled for each patient | 236 | 264 | 270 | 262 | 195 | 192 |
| Avg # sessions with neither primary or secondary therapist | 48% | 28% | 37% | 16% | 37% | 18% |

Figure 1 depicts the number of moderate constraint violations and the objective function value over time once the program is set to run for a particular, but typical, week. Stage One was complete in two seconds. Stage Two took 20 seconds to perform eight iterations during which the number of moderate constraint violations was reduced from five to three and the objective function from 294,190 to 285,190. In Stage Three 23,253 iterations were performed which further reduced the number of moderate constraint violations to one and the objective function to 163,280. Approximately 50,000 unique solutions are checked in the 1.86 seconds that it takes, on average, to compute a Stage Three iteration. Note that 92% of the reduction in objective function is achieved within the first two hours.

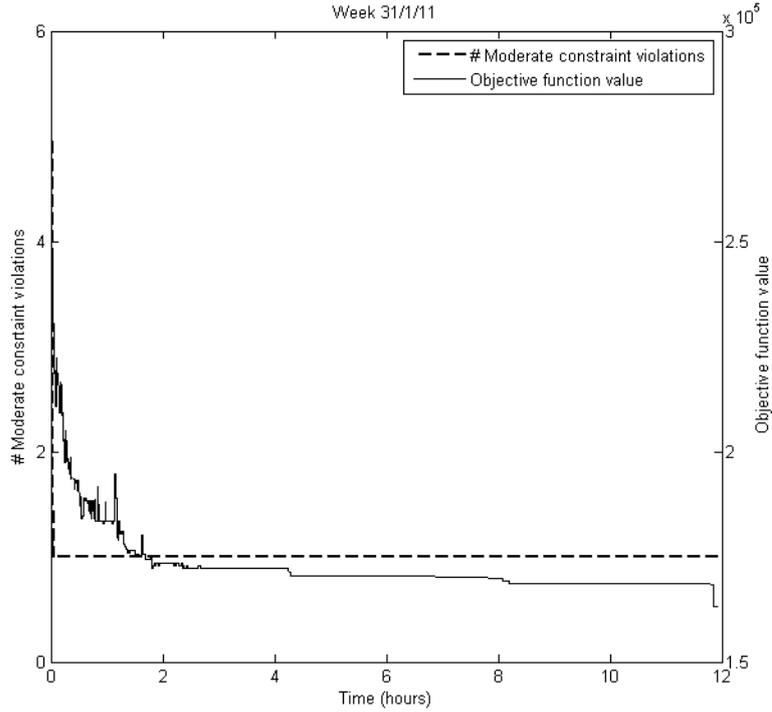


Figure 1 Moderate constraint violations and objective function value over time

3. The queuing system

Having outlined the scheduling procedure, the queuing process for patients is now considered. Arrivals at the system are random, that is, exponentially distributed inter-arrival times (with arrival rate given by $\lambda > 0$). The service time distribution is, in essence, a $k \in \mathbb{N} \geq 2$ phase (general) phase-type distribution⁷ made up of a $k-1$ phase Coxian phase-type distribution⁸ (with service phase rates $\mu_i > 0$ for $i = 1, 2, \dots, k-1$ and exit probabilities $0 \leq \alpha_i \leq 1$ for $i = 1, 2, \dots, k-2$) followed, in series, by an exponential distribution (with service rate $\mu_k > 0$). There are $r \in \mathbb{N}^+$ servers and a total capacity of $N \in \mathbb{N} \geq r$. Hence the buffer size is $N-r$. In Kendall's notation this queuing system is expressed $M | PH_{k-1,1} | r | N$.

3.1 Example

Consider the simple queuing system with $k = 3$, $r = 2$, $N = 3$ depicted in Figure 2.

⁷ See Section 2 of Asmussen et al, 1996 for a specification of phase-type distributions.

⁸ See Section 2 of Marshall & Zenga, 2010 for a specification of the Coxian distribution.

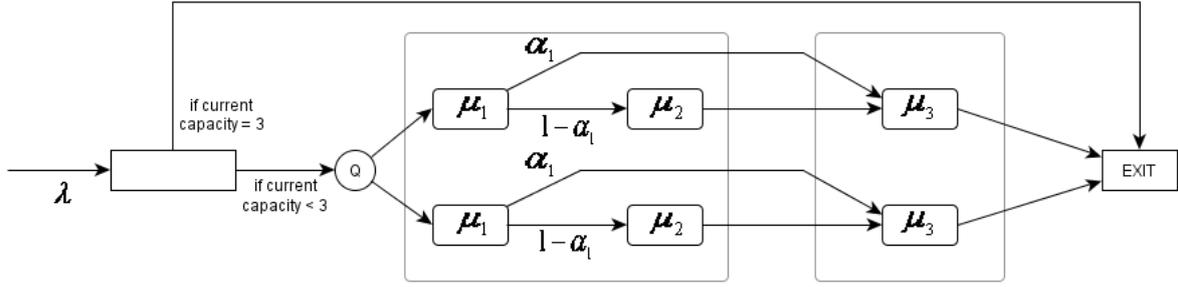


Figure 2 The $M | PH_{2,1} | 2 | 3$ queuing system

Since the time spent in each service phase is exponentially distributed the Markov property holds and so such systems can be represented as a continuous-time Markov chain. The state of the system $s \in S$ is defined by the $k + 1$ -tuple $\langle n, n_1, n_2, \dots, n_k \rangle$ where S is the state space, n is the number in system (i.e. service and queue) and n_i is the total number in service phase $i = 1, 2, \dots, k$. Figure 3 describes the Markov chain for the example.

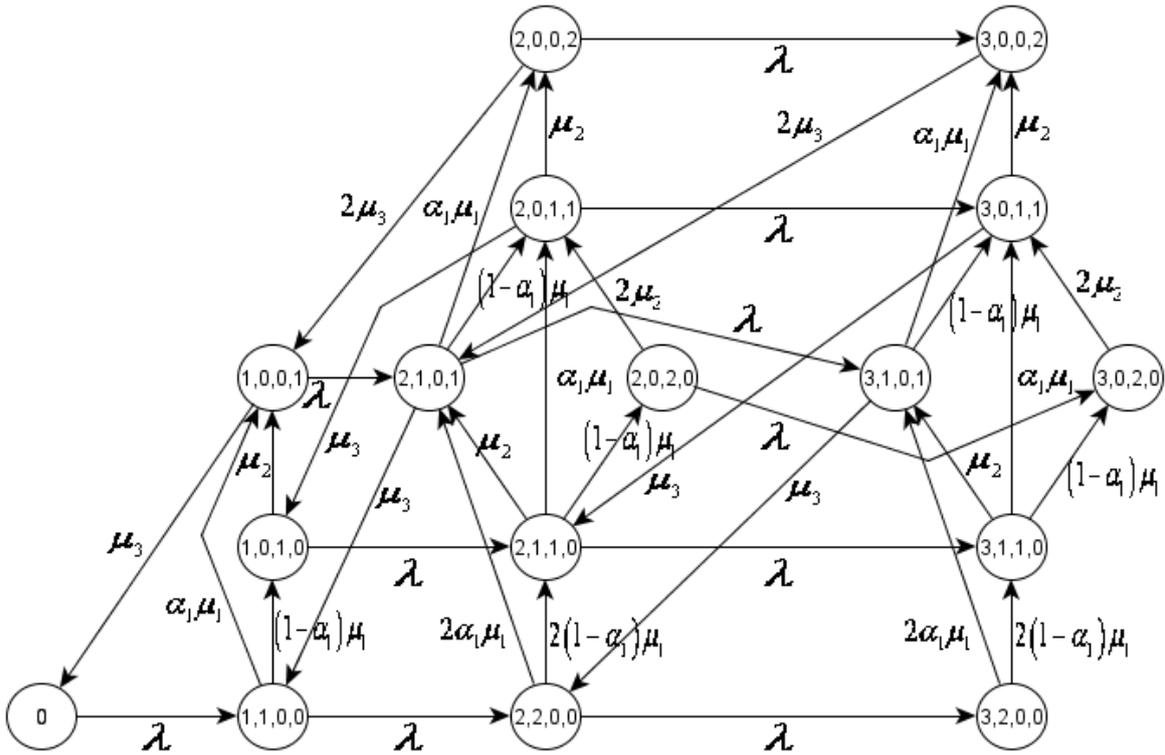


Figure 3 Markov chain representation of the $M | PH_{2,1} | 2 | 3$ queuing system

Note this is a generalisation of the birth-death process synonymous with the $M | M | 1$ and $M | M | r$ systems. For steady-state, the rate into each state must equal the rate out. For example, the inward rate to state $\langle 2, 1, 0, 1 \rangle$ is $\lambda P_{1,0,0,1} + 2\alpha_1\mu_1 P_{2,2,0,0} + \mu_2 P_{2,1,1,0} + 2\mu_3 P_{3,0,0,2}$

If $n > r$ then the state space is equivalent to that of $n = r$ (i.e. $S_{n>r} = S_r$) except, of course, for the value of n . This is because arriving customers join the queue and do not affect the number in each service phase. Thus

$$\eta_n = \binom{r+k-1}{r} \quad (8)$$

and so the total number of states is given by

$$\eta_{total} = \sum_{n=0}^r \binom{n+k-1}{n} + (N-r) \binom{r+k-1}{r} \quad (9)$$

3.3 Formulae for the population of the transition matrix

The general idea is to derive formulae to calculate the position and value of non-zero elements within each non-null sub-matrix. This is done using the following formula:

$$\tau_{e+r_e, e+c_e} = z_e \quad (10)$$

Here $\tau_{i,j}$ denotes the value of the element on the i -th row and j -th column of the sub-matrix in question. The $r_e \in \mathbb{N}^*$ and $c_e \in \mathbb{N}^*$ denote the row and column shift for each non-zero element $e \in \mathbb{N}^+$ within the sub-matrix.

3.3.1 For $n \leq r$

Table 3 contains formulae to populate non-zero elements in the Left, Middle, and Right sub-matrices $t_{i,j}$ with $i, j \leq r$, i.e. those contained in the ‘square’ in eqn. (6). The formulae were deduced by observing the patterns and trends associated with transition matrices produced with various k and r .

As an example, consider the column shift of the Right sub-matrix $t_{i,i+1}$ with $k=2$ and $i \geq 0$. The non-zero elements of this matrix relate to service completion of any of the $i+1$ customers in the system. Such an event can only occur when at least one customer is in the final phase of service. The ‘from’ state (column) must therefore have $n=i+1$ and $n_2 \geq 1$. It can be seen from Table 2 that if $i=0$ this corresponds to the second (underlined) entry in the segment with $k=2, n=1$, i.e. $e = \{1\}$ and $c_1 = 1$. Similarly, if $i=1$ then this corresponds to the second

and third entries of the segment with $k=2, n=2$, i.e. $e = \{1, 2\}$ and $c_1 = c_2 = 1$. The value of c_e can be deduced as the value contained in the e -th set of curly brackets in the expansion of

$$\sum_{l_1=1}^{i+1} \left\{ \sum_{p=1}^1 \binom{l_1 - p}{1 - p} \right\} \quad (11)$$

If $k=3$ then the ‘from’ state (column) must have $n_3 \geq 1$. If $i=0$, then $e = \{1\}$ and $c_1 = 2$. If $i=1$, then $e = \{1, 2, 3\}$ and $c_1 = 2, c_2 = c_3 = 3$ (see eqn. (4)). If $i=2$, then $e = \{1, 2, \dots, 6\}$ and $c_1 = 2, c_2 = c_3 = 3, c_4 = c_5 = c_6 = 4$. It becomes apparent that the value of c_e is contained in the e -th set of curly brackets in the expansion of

$$\sum_{l_1=1}^{i+1} \sum_{l_2=1}^{l_1} \left\{ \sum_{p=1}^2 \binom{l_1 - p + 1}{2 - p} \right\} \quad (12)$$

For $k=4$, this becomes

$$\sum_{l_1=1}^{i+1} \sum_{l_2=1}^{l_1} \sum_{l_3=1}^{l_2} \left\{ \sum_{p=1}^3 \binom{l_1 - p + 2}{3 - p} \right\} \quad (13)$$

It is found that for general $k \geq 2$ the value of c_e is contained in the e -th set of curly brackets in the expansion of

$$\sum_{l_1=1}^{i+1} \sum_{l_2=1}^{l_1} \dots \sum_{l_{k-1}=1}^{l_{k-2}} \left\{ \sum_{p=1}^{k-1} \binom{l_p - p + k - 2}{k - p - 1} \right\} \quad (14)$$

where $e = 1, 2, \dots, \eta_i$.

3.3.2 For $n > r$

The Left, Middle, and Right sub-matrices are all square and of dimension given by eqn. (8). The Left sub-matrices are defined on $i = r+1, r+2, \dots, N$ and are diagonal with elements equal to λ . The Middle sub-matrices are defined on $i = r+1, r+2, \dots, N$ and are equivalent to $t_{r,r}$. The Right sub-matrices are defined on $i = r, r+1, \dots, N-1$ and equal $\begin{bmatrix} t_{r-1,r} \\ \ddot{Z} \end{bmatrix}$ where \ddot{Z} is a null matrix with $\eta_r - \eta_{r-1}$ rows and η_r columns.

3.3.3 Diagonal entries

The rates of transition from a particular state to another have been considered in this subsection thus far. What has not been studied, however, is the rate at which a current state is occupied. These negative-valued rates populate the diagonal entries of the transition matrix (and Middle sub-matrices). Since row totals must equal zero in a transition matrix the diagonal entries of Q^T are populated by subtracting column totals from zero.

Table 3 Formulae to populate non-zero, non-negative entries of non-null sub-matrices with number in system no greater than number of servers

| Sub-matrix | Support | Event | Variable | Support | Value is e-th term in the expansion of |
|---------------------|------------------------|-----------------------------------------------------------------------------------------|-------------------------|-------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $t_{i,i-1}$ (Left) | $i = 1, 2, \dots, r$ | Arrival | r_e c_e z_e | $e = 1, 2, \dots, \eta_{i-1}$ $e = 1, 2, \dots, \eta_{i-1}$ $e = 1, 2, \dots, \eta_{i-1}$ | N/A N/A $\sum_{l=1}^{\eta_{i-1}} \{\lambda\}$ |
| $t_{i,i}$ (Middle) | $i = 0, 1, \dots, r$ | Service phase completion from j-th to j+1-th phase ($j = 1, 2, \dots, k-1$) | r_e c_e z_e | $e = 1, 2, \dots, \eta_{i-1}$ $e = 1, 2, \dots, \eta_{i-1}$ $e = 1, 2, \dots, \eta_{i-1}$ | $\sum_{l_1=1}^i \sum_{l_2=1}^{l_1} \dots \sum_{l_{j-1}=1}^{l_{j-2}} \sum_{l_j=1}^{l_{j-1}+k-j-2} \left\{ \sum_{p=1}^j \binom{l_p - p + k - 2}{k - p - 1} \right\}$ $\sum_{l_1=1}^i \sum_{l_2=1}^{l_1} \dots \sum_{l_{j-1}=1}^{l_{j-2}} \sum_{l_j=1}^{k-j} \left\{ \sum_{p=1}^{j-1} \binom{l_p - p + k - 2}{k - p - 1} \right\}$ $\sum_{l_1=1}^i \sum_{l_2=1}^{l_1} \dots \sum_{l_j=1}^{l_{j-1}+k-j-2} \sum_{l_{j+1}=1}^{l_j} \left\{ (l_{j-1} - l_j + 1)(1 - \alpha_j) \mu_j \right\}$ (n.b. $\alpha_{k-1} = 0$) |
| | | Service phase completion from j-th to k-th phase ($j = 1, 2, \dots, k-2$) | r_e c_e z_e | $e = 1, 2, \dots, \eta_{i-1}$ $e = 1, 2, \dots, \eta_{i-1}$ $e = 1, 2, \dots, \eta_{i-1}$ | $\sum_{l_1=1}^i \sum_{l_2=1}^{l_1} \dots \sum_{l_{j-1}=1}^{l_{j-2}} \sum_{l_j=1}^{l_{j-1}+k-j-1} \left\{ \sum_{p=1}^{k-1} \binom{l_p - p + k - 2}{k - p - 1} \right\}$ $\sum_{l_1=1}^i \sum_{l_2=1}^{l_1} \dots \sum_{l_{j-1}=1}^{l_{j-2}} \sum_{l_j=1}^{k-j} \left\{ \sum_{p=1}^{j-1} \binom{l_p - p + k - 2}{k - p - 1} \right\}$ $\sum_{l_1=1}^i \sum_{l_2=1}^{l_1} \dots \sum_{l_j=1}^{l_{j-1}+k-j-2} \sum_{l_{j+1}=1}^{l_j} \left\{ (l_{j-1} - l_j + 1) \alpha_j \mu_j \right\}$ |
| $t_{i,i+1}$ (Right) | $i = 0, 1, \dots, r-1$ | Service phase completion from k-th phase | r_e c_e z_e | $e = 1, 2, \dots, \eta_i$ $e = 1, 2, \dots, \eta_i$ $e = 1, 2, \dots, \eta_i$ | N/A $\sum_{l_1=1}^{i+1} \sum_{l_2=1}^{l_1} \dots \sum_{l_{k-1}=1}^{l_{k-2}} \left\{ \sum_{p=1}^{k-1} \binom{l_p - p + k - 2}{k - p - 1} \right\}$ $\sum_{l_1=1}^{i+1} \sum_{l_2=1}^{l_1} \dots \sum_{l_{k-1}=1}^{l_{k-2}} \left\{ (l_{k-1}) \mu_k \right\}$ |

3.4 Balking and reneging

Balking is a concept in which arriving customers decide not to join the queue. It is incorporated into the queuing system to model the dissuading effect that long queue lengths or waiting times may have on arrival rates. The arrival rate given n in system is defined by

$$\lambda_n = \lambda(1 - Pb_n) \quad (15)$$

where Pb_n is the probability of an arriving customer balking. Clearly, this probability should be zero when there is at least one service channel available and one when capacity is reached (thus balking is technically part of any finite-buffer system). In between, the probability of balking decreases polynomially (with exponent $\delta > 0$), i.e.

$$Pb_n = \begin{cases} 0 & 0 \leq n < r \\ \left(\frac{n-r+1}{N-r}\right)^\delta & r \leq n \leq N-1 \\ 1 & n \geq N \end{cases} \quad (16)$$

Reneging is a concept in which arriving customers join the queue, but leave before entering service. Here the probability of a queuing customer reneging given $n > r$ in system is defined by

$$Pr_n = \gamma \left(\frac{n-r}{N-r}\right)^\theta \quad (17)$$

with exponent $\theta > 0$ and scaling factor $\gamma > 0$. Whilst balking affects the Left sub-matrices $t_{i,i-1}$ for $i = r+1, r+2, \dots, N$, reneging affects the Right sub-matrices $t_{i,i+1}$ for $i = r, r+1, \dots, N-1$. More specifically it affects the diagonal entries of these sub-matrices which represent the rate of transition from state $\langle n, n_1, \dots, n_k \rangle$ to $\langle n-1, n_1, \dots, n_k \rangle$ for $n > r$. The actual rates $\tau_{e,e}$ for $e = 1, \dots, \eta_r$ are determined from the sum of all other non-negative column values in Q^T (i.e. the other outward rates) and the Pr_n such that

$$\frac{\tau_{e,e}}{\tau_{e,e} + \text{column total}} = Pr_n \quad (18)$$

3.5 Performance measures

Once the steady-state probabilities, P_s , have been found, some performance measures of the system can be deduced (Table 4).

Table 4 Performance measures of the $M | PH_{k-1,1} | r | N$ system with balking and renegeing

| | | |
|-------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| \bar{P}_n | Probability of m in system | $\sum_{\forall s \in S_{n=m}} P_s$ |
| P_R | Probability that an arrival is rejected | \bar{P}_N |
| P_W | Probability that an arrival must wait | $\sum_{n=r}^N \bar{P}_n$ |
| $\tilde{\lambda}$ | Mean effective arrival rate | $\sum_{n=0}^N \lambda_n \bar{P}_n$ |
| π | Mean occupancy | $\sum_{n=0}^r n \bar{P}_n + r \sum_{n=r+1}^N \bar{P}_n$ |
| $\dot{\pi}$ | Server utilisation | π/r |
| $E[\Upsilon]$ | Expected service time | $\mu_1^{-1} + \mu_k^{-1} + \sum_{i=2}^{k-1} \mu_i^{-1} \sum_{j=1}^{i-1} (1 - \alpha_j)$ |
| Tp | Throughput | $\pi/E[\Upsilon]$ |
| $E[L]$ | Expected number in system | $\sum_{n=0}^N n \bar{P}_n$ |
| $E[T]$ | Expected time in system (through Little's Law) | $\tilde{\lambda}^{-1} E[L]$ |
| $E[W]$ | Expected waiting time (before service or renege) | $E[T] - E[\Upsilon]$ |
| $E[L_q]$ | Expected number in queue | $\tilde{\lambda} E[W]$ |
| $\tilde{P}b$ | Mean probability of balking | $\sum_{n=0}^N P b_n \bar{P}_n$ |
| $\tilde{P}r$ | Mean probability of renegeing (ψ_s is rate of renegeing from state s) | $\tilde{\lambda}^{-1} \sum_{m=r+1}^N \sum_{\forall s \in S_{n=m}} \psi_s P_s$ |

It is important to note that since the system is in steady-state, the mean effective arrival rate must equal the mean departure rate, i.e.

$$\tilde{\lambda} = \pi E[\Upsilon]^{-1} + \tilde{\lambda} \tilde{P}r \quad (19)$$

3.6 Computer automation

A computer program in Maple has been set up to firstly populate the transition matrix (using the formulae of Section 3.3); then to solve it (using *LinearSolve*); and finally to output performance measures (Section 3.5) in addition to providing some graphical outputs. Before running the program the user must input numeric values for k , r , N , and numeric or symbolic values for λ , μ_1, \dots, μ_k , $\alpha_1, \dots, \alpha_{k-2}$ and δ , γ , θ if balking and renegeing are included (both optional).

The solution of the linear system is by far the most time-consuming of the three tasks. The time taken is dependent on the dimension and density of the transition matrix and the precision used in operations. For example, on a standard desktop computer, a $M | PH_{1,1} | 10 | 20$ system has 176 states and takes 12 seconds to solve analytically for numeric results whilst a $M | PH_{1,1} | 20 | 50$ has 861 states and requires almost ten minutes. However, a $M | PH_{2,1} | 10 | 15$ has fewer states (616) but takes slightly more time due to a denser¹¹ transition matrix.

4. Data

4.1 Arrivals dataset

Data is only available for the sixteen months from January 2010 until April 2011. In this time the mean arrival rate is $\tilde{\lambda} = 0.36364$ referrals¹² per day. The assumption of random arrivals is justified (Figure 4). The average queue length is $E[L_q] = 15.83$ with a standard deviation of 3.43, a minimum of nine and a maximum of 25. The average waiting time is bounded below (since some observations are censored) at $E[W] > 38$ days. Only one quarter of referrals have resulted in admission to Rookwood Hospital. No evidence has been found to suggest that the probability of renegeing is based on queue size or duration of time in queue.

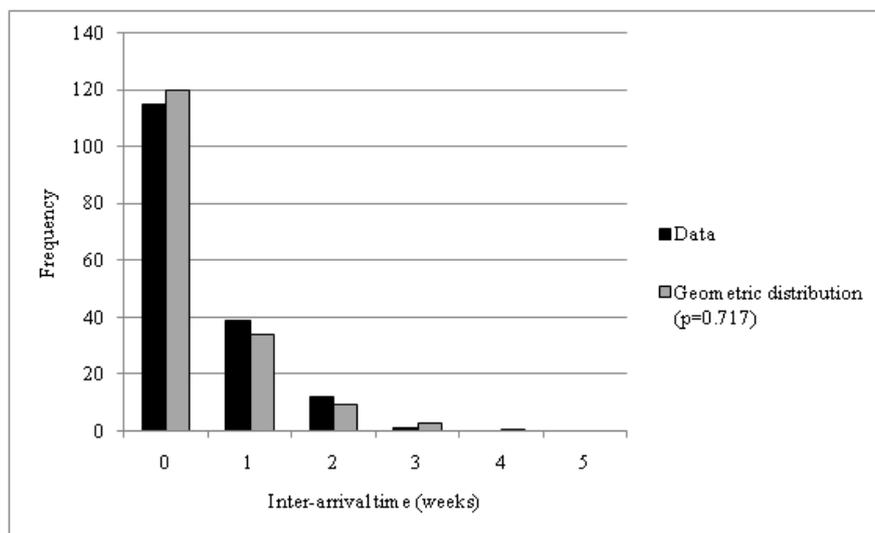


Figure 4 Frequency chart for empirical data and the geometric distribution

¹¹ Since there are more service phases for which transitions can occur.

¹² Referred patients are added to the waiting list where they remain until either admitted or renegeed.

4.2 Service dataset

Data is available for the eight years from 2003 until 2011¹³. The mean LOS of the 358 episodes for which data is available is $\hat{L\hat{O}S} = E[\Upsilon] = 149.3$ days, with a coefficient of variation of 1.018. Each episode contains data on admission age, gender, local health board, primary diagnosis, admission source, admission method, discharge destination. Typical occupancy is about 21 beds in recent years. Figure 5 depicts the relationship between treatment intensity and active LOS for the 89 episodes that have such information.

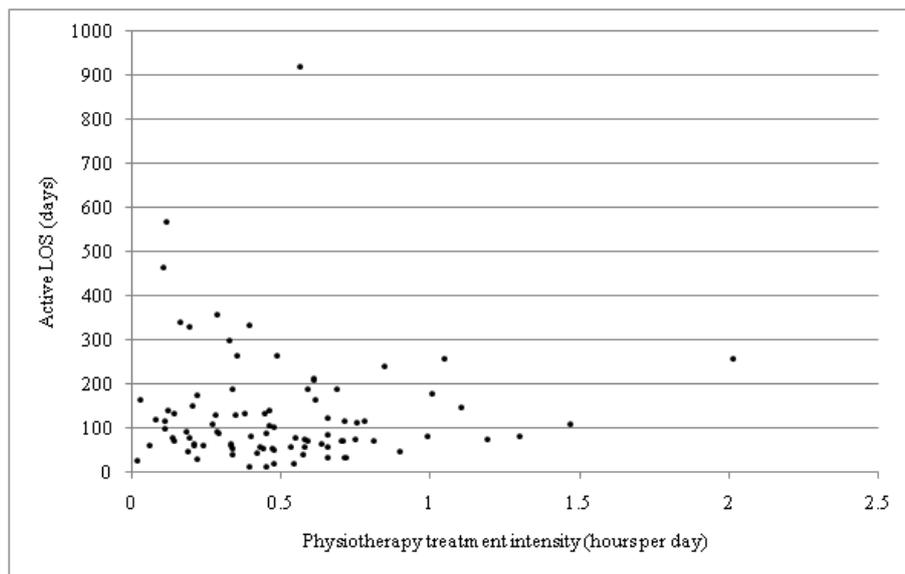


Figure 5 Plot of physiotherapy treatment intensity against active LOS

5. Model construction

The $M | M | 21$ system serves as a simple queuing model of the unit at Rookwood Hospital. A more representative model is attained by partitioning LOS into its active (admission date until readiness date) and blocked (readiness date until discharge date) components. If the former is modelled by the Coxian distribution and the latter by the exponential then the $M | PH_{k-1,1} | r | N$ system of Section 3 can be used. However, this does not take into account patient attributes that affect LOS. These are included by allowing the service rates and number of phases in the Coxian distribution to differ from server to server. It has been found (Wood, 2011) that, for fixed k , the number of states is given by

¹³ Earlier records are available but are not used in order to promote time-heterogeneity of data – the current superintendent physiotherapist took up his post in 2003.

$$\eta_{total} = \sum_{i=0}^r \binom{r}{i} k^i + (N-r)k^r \quad (20)$$

So for a system with $k = 3$, $r = 20$ and $N = 50$, the dimension of the transition matrix would be 2,229,556 compared with just 1,606 with homogeneous servers (i.e. same number of phases and service rates in each service channel). Clearly this is too large for analysis, and so an alternative is favoured.

Here, a small number, p , of patient groups are formed based on the values of the most significant patient attributes that return a low variance of active LOS within each group (Stage One). The homogenous-server queuing system of Section 3 (i.e. $M | PH_{k_i-1,1} | r_i | N_i$ for $i = 1, \dots, p$ with $r = \sum r_i = 21$) is thereafter used to model the LOS of each patient group (Stage Four). But the active service phase rates are dependent on treatment intensity. The automated scheduling program is used to produce group average values for physiotherapy intensity (Stage Two). This is done using typical values of the input variables (Section 2) since, for model construction, a representation of reality is sought. The resulting treatment intensity levels must equal the empirical averages for each patient group. Graphs plotting group treatment intensity against group active LOS (Stage Three) are then used to obtain average values of active LOS for each patient group. For model construction these must equal their empirical counterparts.

5.1 Stage One: Patient groups

Classification and Regression Tree (CART) analysis is used to produce a binary decision tree by performing, at each iteration, a two-way split on a particular variable based on the largest reduction in variance. In this case, the target variable is active LOS and the predictor variables are listed in Section 4.2. The program Treeworks (Harper & Leite, 2008) is used to produce four mutually exclusive groups (Table 5) with a 21% reduction in variance. The appendix contains summary statistics for LOS based on the three patient attributes that were branched on.

Table 5 Specification of the four patient groups

| Group | Discharge destination | Admission Source | Diagnosis |
|-------|---------------------------------------------------------|-------------------------------------------------|----------------------------------------|
| 1 | Usual residence, Temporary residence | Cardiff University hospital, Usual residence | (any) |
| 2 | Usual residence, Temporary residence | Other NHS hospital | (any) |
| 3 | Internal transfer, Other Trust, Non-NHS nursing home | (any) | GBS, Anoxia, MS, Other neuropathies |
| 4 | Internal transfer, Other Trust, Non-NHS nursing home | (any) | ABI, TBI, Other |

If n_i and $\hat{L}OS_i$ denote the sample size and mean LOS of group i , then $n_1 = 183$, $n_2 = 40$, $n_3 = 41$, $n_4 = 94$ and $\hat{L}OS_1 = 86.27$, $\hat{L}OS_2 = 162.02$, $\hat{L}OS_3 = 175.49$, $\hat{L}OS_4 = 255.21$. The proportion of service channels assigned to the queuing system of each group is set equal to the proportion of bed days occupied by patients of that group, i.e.

$$r_i = \frac{n_i \cdot \hat{L}OS_i}{\sum_{i=1}^p n_i \cdot \hat{L}OS_i} \cdot r \quad (21)$$

Since $r_i \in \square^+$ they are rounded such that $r_1 = 6$, $r_2 = r_3 = 3$, $r_4 = 9$.

5.2 Stage Two: Treatment intensity

The automated scheduling program described in Section 2 is amended to deal with groups of patients instead of individuals. Here, typical values of patient demand, details and priority are entered for each group (not for each patient). On the supply-side, a typical skill-mix is entered alongside typical working hours and availabilities. Once the schedule is run, a number of performance measures are output – the most important being ‘group average clinical minutes received’. After five runs of the program the average intensity of physiotherapy for each group is $\hat{INT}_1 = 0.4333$, $\hat{INT}_2 = 0.3905$, $\hat{INT}_3 = 0.6262$, $\hat{INT}_4 = 0.5214$ hours per day. These are all very close to the actual/empirical group averages obtained through the Service dataset.

5.3 Stage Three: Service rates

Scatter plots similar to Figure 5 are produced for each patient group. A line of best fit of active LOS against treatment intensity is then fitted to each graph. If \hat{n}_i denotes the number of episodes of each group i that have data on physiotherapy intensity and date ready for discharge then $\hat{n}_1 = 46$, $\hat{n}_2 = 13$, $\hat{n}_3 = 7$, $\hat{n}_4 = 23$. Let x_j and y_j denote the physiotherapy intensity and active LOS of observation $j \in \{1, \dots, \hat{n}_i\}$ for group i . The line of best fit given by the equation

$$y_j = \beta_1 + \frac{\beta_2}{x_j} \quad (22)$$

is fitted to the plot by the method of least squares by minimising

$$\sum_{\forall j} \left(y_j - \beta_1 - \frac{\beta_2}{x_j} \right)^2 \quad (23)$$

The constraint

$$L\hat{O}S_i^{active} = \beta_1 + \frac{\beta_2}{I\hat{N}T_i} \quad (24)$$

is added to ensure that the average level of physiotherapy intensity corresponds to the average active LOS for each group (obtained in Section 5.4).

5.4 Stage Four: Queuing system

Coxian distributions are fitted to active LOS by maximum likelihood through the Nelder-Mead 'simplex' algorithm in Matlab. The approximations are displayed in Figure 6.

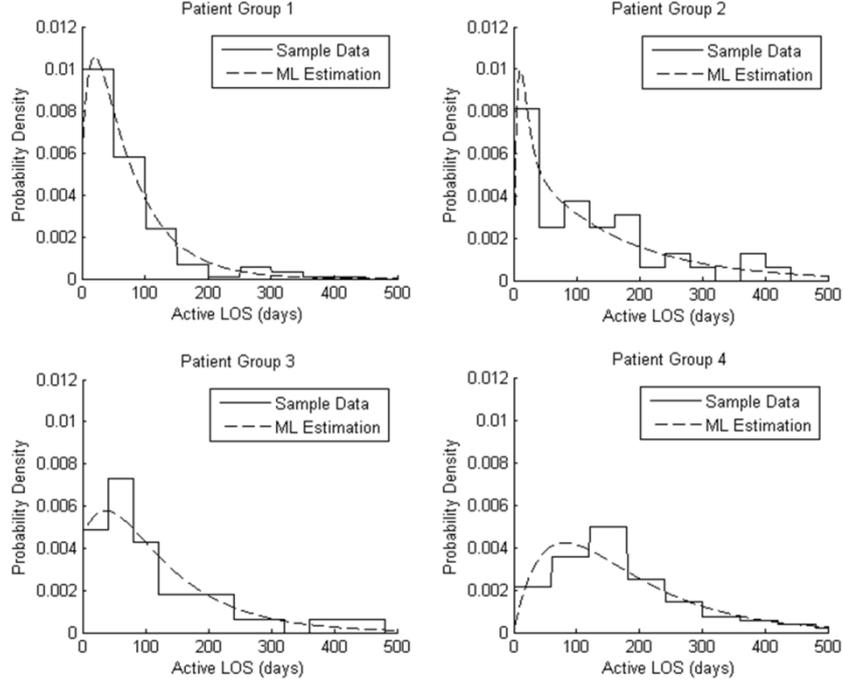


Figure 6 Coxian approximations to active LOS for each patient group

Here, $k_1^{active} = k_4^{active} = 2$ and $k_2^{active} = k_3^{active} = 3$ where k_i^{active} is the number of phases in the Coxian distribution of patient group i . The Bayesian Information Criterion and Akaike Information Criterion are used to balance goodness of fit with tractability (as more phases implies higher transition matrix dimension). The mean of each distribution is

$\hat{LOS}_1^{active} = 72.22$, $\hat{LOS}_2^{active} = 136.55$, $\hat{LOS}_3^{active} = 120.87$, $\hat{LOS}_4^{active} = 177.53$. These are approximately equal to their empirical counterparts. The exponential distribution is fitted to blocked LOS for each group by taking the rate parameter as the reciprocal of the empirical averages, $\hat{LOS}_1^{blocked} = 13.63$, $\hat{LOS}_2^{blocked} = 26.35$, $\hat{LOS}_3^{blocked} = 53.12$, $\hat{LOS}_4^{blocked} = 77.09$.

The buffer size of each system is calculated as the proportion of beds allocated to each group multiplied by total buffer size, i.e.

$$N_i = r_i + \frac{r_i}{r} N \quad (25)$$

The total buffer size is set at 28 (maximal queue size from Arrivals dataset is 25) and so $N_1 = 14$, $N_2 = N_3 = 7$, $N_4 = 21$.

The final step in parameterising the model is to estimate the four remaining parameters, λ , δ , γ , θ . Estimations of these cannot be found directly and so they are determined by trial

and error such that the performance measures of the resulting model are equivalent to some target values. Unfortunately, the results of Section 4.1 cannot be used as targets since the timescales of the Arrivals and Service datasets do not correspond. Moreover, it is clinical opinion that LOS has been noticeably larger in the sixteen months of the Arrivals dataset than in years past. However, there is no reason to suppose that the effective arrival rate has changed over time. The targets for the effective arrival rates into each of the four systems are deduced. They cannot be derived exactly through the Arrivals dataset since the patient attributes (listed in Section 4.2) are not recorded for each referral. Instead the proportion of total arrivals (0.36364) to each patient group is set equal to their proportion of total throughput, i.e.

$$\tilde{\lambda}_i = \frac{\frac{r_i}{L\hat{O}S_i}}{\sum_{t=1}^4 \frac{r_t}{L\hat{O}S_t}} \cdot \tilde{\lambda} \quad (26)$$

Thus, $\tilde{\lambda}_1 = 0.18010$, $\tilde{\lambda}_2 = 0.04795$, $\tilde{\lambda}_3 = 0.04427$, $\tilde{\lambda}_4 = 0.09132$. Appropriate targets for the mean probabilities of renegeing are now deduced using the mean service time from the Service dataset. Since the mean occupancy of each system is bounded above by the number of servers then, from eqn. (19),

$$\tilde{\lambda}_i < r_i E[Y_i]^{-1} + \tilde{\lambda}_i \tilde{P}r_i \quad (27)$$

By substitution of eqn. (26) into eqn. (27), and noting $E[Y_i] = L\hat{O}S_i$, returns

$$\tilde{P}r_i > 1 - \frac{\sum r_t E[Y_t]^{-1}}{\tilde{\lambda}} = 0.614 \quad (28)$$

The parameters λ , δ , γ , θ are now varied such that the $\tilde{\lambda}_i$ and $\tilde{P}r_i$ are approximately equal to these targets. Steady-state probabilities are deduced for the holistic system by the multiplication of those for each of the four constituent systems (due to independence). Thus,

$$\begin{aligned} & P\left(\langle n_1, n_{1,1}, \dots, n_{1,r_1} \rangle \cup \langle n_2, n_{2,1}, \dots, n_{2,r_2} \rangle \cup \langle n_3, n_{3,1}, \dots, n_{3,r_3} \rangle \cup \langle n_4, n_{4,1}, \dots, n_{4,r_4} \rangle\right) \\ &= \prod_{i=1}^4 P\left(\langle n_i, n_{i,1}, \dots, n_{i,r_i} \rangle\right) \end{aligned} \quad (29)$$

Figure 7 depicts the probability density for the number in system.

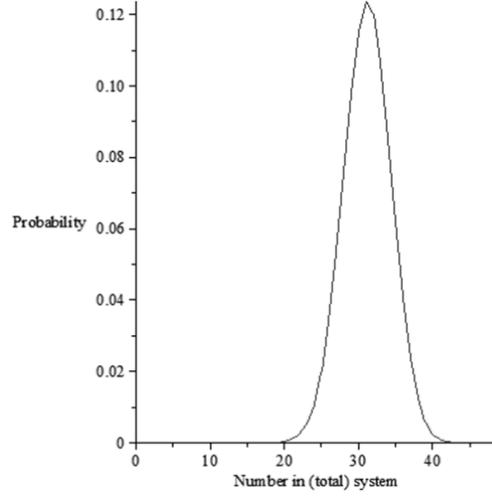


Figure 7 Graph of probability density for number in system

The performance measures for the holistic system are deduced as follows. If

$P \in \{L, L_q, \tilde{\lambda}, Tp, \pi\}$ then

$$E[P_{sys}] = \sum_{i=1}^4 E[P_i] \quad (30)$$

If $P \in \{W, T, \tilde{Pr}\}$ then each measure is weighted by their throughput, i.e.

$$E[P_{sys}] = \left(\sum_{i=1}^4 Tp_i \right)^{-1} \sum_{i=1}^4 Tp_i E[P_i] \quad (31)$$

The value of these performance measures are displayed in Table 6.

Table 6 Performance measures of the model

| | | |
|-------------------------|--------|-----------------------|
| $\tilde{\lambda}_{sys}$ | 0.3629 | referrals/day |
| π_{sys} | 20.76 | beds |
| \tilde{Pr}_{sys} | 0.617 | |
| Tp_{sys} | 50.79 | episodes/year |
| $E[L_{sys}]$ | 31.16 | referrals or patients |
| $E[L_{q,sys}]$ | 10.4 | referrals |
| $E[T_{sys}]$ | 85.89 | days |
| $E[W_{sys}]$ | 28.66 | days |

Thus there is a 0.2% difference between the model and target mean effective arrival rates and a 0.5% difference between the model and target mean probabilities of renegeing. In the absence of other numerical targets, clinicians have examined graphical outputs and other performance measures in order to confirm an accurate representation of the unit. The annual running cost of the unit (with a bed-cost per day of £480) is estimated through the formulae

$$\pi_{sys} \times £480 \times 365 = £3.637m \quad (32)$$

6. Hypothetical scenarios

Many what-if type scenarios have been considered in this study involving changes to the number of beds, composition of therapy, and skill-mix of staff. The effects of an ageing population have also been studied by creating two patient groups based on age and increasing the arrival rate into the one that contains older patients. Two hypothetical scenarios are forthwith examined in further detail.

6.1 Reduction in blocked LOS

The scenario concerning a 50% reduction in blocked LOS can be evaluated by making a change only to Stage Four. Here the blocked service rates for each system are increased two-fold. This yields a mean occupancy of $\pi = 20.65$, a mean probability of renegeing of $\tilde{Pr} = 0.576$, an annual throughput of $Tp = 56.68$, an average wait of $E[W] = 25.71$ and an expected queue length of $E[L_q] = 9.42$. Using eqn. (32) returns an annual running cost of £3.618m. Thus the average cost per patient episode has decreased from £72k to £64k.

If a 100% reduction can be achieved then $\pi = 20.4$, $\tilde{Pr} = 0.452$, $Tp = 59.54$, $E[W] = 22.33$ and $E[L_q] = 8.21$. Annual costs are reduced to £3.574m (£60k per episode).

6.2 Workforce enlargement

This hypothetical scenario involves making a change to the automated scheduling program (Stage Two) since the size (and skill-mix) of the workforce affects group average treatment intensity values. In this example the number of FTEs is increased by one third (whilst retaining a proportionate skill-mix). After five runs of the program the average intensity of physiotherapy for each group is $\hat{INT}_1 = 0.6$, $\hat{INT}_2 = 0.4929$, $\hat{INT}_3 = 0.7857$, $\hat{INT}_4 = 0.7143$.

Using these and eqn. (22) for each system yields $L\hat{O}S_1^{active} = 70.01$, $L\hat{O}S_2^{active} = 125.61$, $L\hat{O}S_3^{active} = 112.44$, $L\hat{O}S_4^{active} = 164.88$ (Stage Three). Finally, the active LOS distributions are amended such that their means are equivalent to these values. This is done by making proportionate adjustments to the service rates whilst holding constant the transition probabilities. This ensures that the underlying shape of the approximating distribution remains the same but the expected values are appropriately scaled. It is done by solving the following equation (based on 'expected service time' in Table 4) for \hat{p}_i :

$$L\hat{O}S_i^{active} = \frac{1}{\hat{p}_i \mu_{i,1}} + \sum_{l=2}^{k_i^{active}} \frac{1}{\hat{p}_i \mu_{i,l}} \sum_{j=1}^{l-1} (1 - \alpha_j) \quad (33)$$

The new active service rates are the product of the old service rates and the respective \hat{p}_i .

Solving the holistic queuing system yields $\pi = 20.73$, $\tilde{P}r = 0.604$, $Tp = 52.78$,

$E[W] = 27.62$ and $E[L_q] = 10.08$. The annual cost is $20.73 \times \text{£}480 \times 365 = \text{£}3.632\text{m}$ in addition to the cost of the extra physiotherapists, $\text{£}77\text{k}$. However, in order to affect active LOS, these changes to the workforce must be made to all other departments at a cost of $\text{£}73\text{k}$. So the total annual cost is $\text{£}3.782\text{m}$ and the cost per episode is just under $\text{£}72\text{k}$.

7. Discussion

A queuing model of a neurological rehabilitation unit has been presented. Two major components of this model, the automated scheduling program and a relevant queuing system, are described separately in Sections 2 and 3. The data is then introduced in the following section and the four-stage model is constructed in Section 5. Two hypothetical scenarios are thereafter evaluated in Section 6. The results of these are now discussed.

Clearly there are significant advantages in reducing the delay to discharge. With a 50% reduction in blocked LOS, there is lower bed occupancy and so annual running costs are $\text{£}19\text{k}$ less. The 12% increase in throughput and 7% decrease in the probability of renegeing mean that more patients are benefitting from a specialist rehabilitation programme. Such patients are less of a burden on social services (Turner-Stokes et al, 2006) and are more likely to re-enter the workforce (Cullen et al, 2007). Referred patients can also expect a 17% shorter wait in the queue. This has physical as well as psychological benefit since the condition of the patient can deteriorate in an inappropriate setting (Cope & Hall, 1982). Decreased waiting

time also has a consequent effect in reducing bed-blocking at the referring facilities upstream where bed costs are much higher.

But how can delays to discharge actually be reduced at Rookwood Hospital? One idea is to inform discharge destinations of the expected discharge readiness dates as early as possible. This should allow plenty of time for them to make the necessary arrangements to receive the patient on time. Accurate estimations of expected LOS can be deduced from a binary decision tree (similar to that described in Section 5.1) using the values of the attributes listed in Section 4.2.

Increasing the size of the workforce affects the active component of LOS. Like the previous scenario this has a positive effect on throughput, renegeing and waiting times but requires additional resources. A rival option to improve throughput is to increase the number of beds whilst holding constant treatment intensity. This would cost £70k more per annum but would increase throughput to 53.78 and decrease the probability of renegeing to 0.595. The cost per episode would be approximately equivalent. The purpose of this study is not to recommend a particular strategy but to lay out the pros and cons of a variety of scenarios that have been suggested by clinicians. Management can then determine whether additional costs are outweighed by improvements in performance.

As would be expected there are limitations associated with such a complex model – many of which relate to the (potentially unknown) extent to which assumptions are satisfied. For example, it is assumed in Stage Two that the amount of treatment scheduled by the automated program is equivalent to the amount received. Due to 'shocks' such as employee absence or patient illness this is not always the case. In Stage Four it is assumed that the queue discipline is first-in first-out, whilst in reality the condition of the patient is also a factor. But perhaps the largest limitation of this project has been the availability of data. In populating the Service dataset it was necessary to strike a balance between the quality and quantity of data. Whilst it was possible to collect more data (from years earlier than 2003), the relevancy of this data would be questionable (since clinical practices change over time).

There is plenty of scope for further work. Patient outcome is part of the trilateral relationship (with treatment intensity and LOS) that has been held constant in this study. That is, changes to treatment intensity affect active LOS such that outcome is unaltered. The ability to have this as a free variable would permit the evaluation of some interesting scenarios. Other studies (such as Turner-Stokes, 2007) suggest that a greater outcome equates to less cost per

unit time post-discharge. Here, the time required to offset the additional costs associated with a higher outcome is studied. According to the relationship, a higher outcome can be achieved through increased LOS or treatment intensity or both. If included in the queuing model, then the consequent effect of such alterations on outcome could be evaluated. If the relationship between outcome and cost per unit time post-discharge is known then a number of valuable what-if questions can be asked. For example, is it cost-effective to provide younger patients with a better outcome? Clearly, the integration of inpatient rehabilitation with related facilities (in this case, downstream) presents some interesting opportunities.

In conclusion, this study has showcased the benefits of academic advancement in operational research; the application of novel techniques to novel settings. Not only have theoretical advancements been made in this work but these have been successfully applied to an important real-life process that has seen little research to date. Mr Jakko Brouwers MSc MCSP, Senior Service Improvement Programme Manager and Superintendent Physiotherapist at Rookwood Hospital, remarks that this work has *'had a huge impact in how resources are utilised'*. On the timetabling of physiotherapy treatment he says that *'automated computer scheduling has created a fairer system for patients'*. Mr Brouwers also confirms the benefits bought about by the holistic model, claiming that it *'has been a real asset in that it has opened the eyes of the operational service managers to the issues regarding patient flow'*. He continues: *'These insights are now used on a regular basis in waiting list management and admissions meetings'*. In summary, Mr Brouwers concludes that *'the investment from the department in support of the research has been well worth it'*.

Acknowledgements

The authors would like to acknowledge the help and support of staff at Rookwood Hospital with particular thanks to Jenny Thomas, Jakko Brouwers and Angelo Lamberti. The authors would also like to thank the referee. This work has been funded by the Engineering and Physical Sciences Research Council (EPSRC) through a Science and Innovation (LANCS) initiative.

Appendix

Table A.1 Basis statistics for LOS categorised by discharge destination

| Discharge Destination: | Usual residence | Internal transfer | Other Trust | Non-NHS nurs. home | Temporary residence |
|---------------------------|-----------------|-------------------|-------------|--------------------|---------------------|
| <i>n</i> | 212 | 61 | 50 | 15 | 6 |
| <i>mean</i> <i>LOS</i> | 100.5 | 253.8 | 208.2 | 206.9 | 114.7 |
| <i>s.d.</i> | 109.1 | 215.4 | 136.6 | 111.5 | 135 |

Table A.2 Basic statistics for LOS categorised by admission source

| Admission Source: | Cardiff University hosp | Other NHS hosp | Usual residence | Temp | Unknown |
|---------------------------|-------------------------|----------------|-----------------|------|---------|
| <i>n</i> | 229 | 78 | 49 | 1 | 1 |
| <i>mean</i> <i>LOS</i> | 146.3 | 200.1 | 73.6 | 352 | 387 |
| <i>s.d.</i> | 148.1 | 172.4 | 85.1 | . | . |

Table A.3 Basic statistics for LOS categorised by primary diagnosis

| Primary diagnosis: | ABI | MS | TBI | Anoxia | GBS | Other | Other neuropathies |
|---------------------------|-------|-------|-------|--------|------|-------|--------------------|
| <i>n</i> | 187 | 44 | 42 | 26 | 22 | 14 | 10 |
| <i>mean</i> <i>LOS</i> | 172.8 | 95.7 | 132.8 | 193.1 | 99 | 104.6 | 83.5 |
| <i>s.d.</i> | 170.3 | 100.9 | 126.8 | 155.1 | 88.3 | 91 | 131.2 |

References

- Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4), 419-441.
- Barnes, M. P. (2003). Principles of neurological rehabilitation. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(4), 3-7.
- Blackerby, W. F. (1990). Intensity of rehabilitation and length of stay. *Brain Injury*, 4(2), 167-173.
- Cifu, D. X., Kreutzer, J. S., Kolakowsky-Hayner, S. A., Marwitz, J. H., & Englander, J. (2003). The relationship between therapy intensity and rehabilitative outcomes after traumatic brain injury: a multicenter analysis¹. *Archives of physical medicine and rehabilitation*, 84(10), 1441-1448.
- Cope, D. N., & Hall, K. (1982). Head injury rehabilitation: benefit of early intervention. *Archives of physical medicine and rehabilitation*, 63(9), 433-437.
- Cullen, N., Chundamala, J., Bayley, M., & Jutai, J. (2007). The efficacy of acquired brain injury rehabilitation. *Brain Injury*, 21(2), 113-132.
- Faddy, M. J., & McClean, S. I. (2005). Markov chain modelling for geriatric patient care. *Methods of information in medicine*, 44(3), 369-373.
- Hanlon, R. E. (1996). Motor learning following unilateral stroke* 1. *Archives of physical medicine and rehabilitation*, 77(8), 811-815.
- Harper, P., & Leite Jr, E. (2008). TreeWorks: Advances in Scalable Decision Trees. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, 3(4), 53-68.
- Heinemann, A. W., Hamilton, B., Linacre, J. M., Wright, B. D., & Granger, C. (1995). Functional status and therapeutic intensity during inpatient rehabilitation. *American journal of physical medicine & rehabilitation*, 74(4), 315-326.
- Marshall, A. H., & Zenga, M. (2010). Experimenting with the Coxian Phase-Type Distribution to Uncover Suitable Fits. *Methodology and Computing in Applied Probability*, 1-16.
- McClean, S., & Millard, P. (2007). Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care? *Journal of the Operational Research Society*, 58(2), 255-261.

- Sandhaug, M., Andelic, N., Vatne, A., Seiler, S., & Mygland, A. (2010). Functional level during sub-acute rehabilitation after traumatic brain injury: Course and predictors of outcome. *Brain Injury*, 24(5), 740-747.
- Slade, A., Tennant, A., & Chamberlain, M. A. (2002). A randomised controlled trial to determine the effect of intensity of therapy upon length of stay in a neurological rehabilitation setting. *Journal of rehabilitation medicine*, 34(6), 260-266.
- Spivack, G., Spettell, C. M., Ellis, D. W., & Ross, S. E. (1992). Effects of intensity of treatment and length of stay on rehabilitation outcomes. *Brain Injury*, 6(5), 419-434.
- Taylor, G. J., McClean, S. I., & Millard, P. H. (1998). Using a continuous time Markov model with Poisson arrivals to describe the movements of geriatric patients. *Applied stochastic models and data analysis*, 14(2), 165-174.
- Teasell, R., Bayona, N., Marshall, S., Cullen, N., Bayley, M., Chundamala, J., et al. (2007). A systematic review of the rehabilitation of moderate to severe acquired brain injuries. *Brain Injury*, 21(2), 107-112.
- Turner, A., Foster, M., & Johnson, S. E. (2002). *Occupational therapy and physical dysfunction: principles, skills, and practice* (5th ed.): Elsevier Health Sciences.
- Turner-Stokes, L. (2003). *Rehabilitation following acquired brain injury: national clinical guidelines*: Royal College of Physicians.
- Turner-Stokes, L. (2007). Cost-efficiency of longer-stay rehabilitation programmes: Can they provide value for money? *Brain Injury*, 21(10), 1015-1021.
- Turner-Stokes, L., Paul, S., & Williams, H. (2006). Efficiency of specialist rehabilitation in reducing dependency and costs of continuing care for adults with complex acquired brain injuries. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(5), 634-639.
- Wood. (2011). *Modelling Activities at a Neurological Rehabilitation Unit*. (PhD thesis, Cardiff University). <http://orca-mwe.cf.ac.uk/35074>.