Modelling capacity along a patient pathway with delays to transfer and discharge

R.M. Wood and B.J. Murch

Modelling and Analytics
Bristol, North Somerset, and South Gloucestershire Clinical Commissioning Group
United Kingdom National Health Service

Abstract

This paper presents a versatile model to estimate capacity requirements along a patient pathway with delays to transfer and discharge caused by blocking after service. Blocking after service is a common property in health systems where patients requiring uninterrupted care can only leave a service when they can be admitted downstream to another. Unlike many studies, the approach appreciates both variability in arrivals and length of stay at each service point and the stochastic queuing dynamics between them. This is crucial to the understanding and identification of blockages which can present a significant source of inefficiency. The problem is framed as a continuous-time Markov chain where patient arrivals, transfers, and discharges are modelled through state transitions, with length of stay approximated by the Erlang distribution. The solution is through simulating movements around this chain by dynamically sampling the next state accessible from the neighbourhood of the current, thus bypassing the need for time-intensive manipulations of equations involving the entire transition matrix. This is packaged in easy-to-use code in free software so as to be readily available to healthcare practitioners. An example of use is illustrated for a stroke pathway reconfiguration where move to a centralised hyper-acute service is assessed.

Keywords*:* Queuing theory; Markov processes; Blocking after service; Capacity planning; Stroke

1. Introduction

In a healthcare setting blocking after service, sometimes referred to as Type 1 blocking, occurs when a patient has received their intended course of treatment but cannot proceed to their downstream service point because there is no availability. It is a problem since the patient continues to consume specialist and costly resources they do not need whilst preventing access to someone who does have a need for them.

There are a large number of patient pathways which involve blocking after service. For some this may be over the whole pathway, such as acquired brain injury where continuing care is required from emergency to acute, rehabilitation, and social care. Here blocking can occur at any of these handover points since the patient is required to always be under someone's care. It can also be manifest over certain parts of a pathway, such as an elective pathway where a patient may wait at home between pre-op consultations but post-op could require inpatient rehabilitation followed by social care support. If there is no social care availability then the patient must remain at the rehabilitation unit, even if they are medically fit for discharge and have expended the need for specialist rehabilitation.

Blocking after service is seldom an isolated problem and blockages located toward the end of the pathway can quickly reverberate upstream causing system-wide congestion. For instance, patients occupying beds at rehabilitation units waiting for social care availability fill up the unit and make it harder to accommodate arrivals from the acute units. In turn, the waiting patients in the acute units make it harder to admit patients from emergency units. This is a major problem in the National Health Service (NHS) of the United Kingdom where there are often reports of patients waiting hours in trolleys in emergency units because there are no hospital beds available. Indeed in 2017 there were an estimated 1.8 million bed days lost due to these delays in transfer of care in England (NHS England, 2018).

In the NHS, with rising demand (Keehan et al, 2016) and fixed real-term budget (NHS England, 2014), alleviating these unnecessary problems is critically important. Recent efforts to change the culture away from thinking *in silos* and toward thinking *system-wide* are a step in the right direction (World Health Organisation, 2015, The Health Foundation, 2016). But any modelling on the ground to support this welcome change in culture remains lacking.

1.1 Objectives

The objective of this paper is to construct a versatile capacity model for a patient pathway containing blocking after service, where the detailed dynamics of and crucially between service points are appreciated. In achieving this it is necessary to set out some modelling requirements, as follows.

A key feature required of the model will be that it distinguishes between the *active* and *blocked* components of service point length of stay (LOS). Active LOS is defined by the duration of time from admission until readiness for transfer or discharge while blocked LOS is the remainder of time until ultimate discharge. The crucial difference between the two is that active LOS should be an empirically-calibrated independent variable while blocked LOS should be a dependent variable; dependent on the availability of downstream service points through the pathway dynamics. Another key feature of the model is that it must be *stochastic*, such that the variability that realistically occurs in the main dynamical components of the pathway can be realised. Only through realising these variabilities can blocking after service be reliably modelled.

The parameters of the model should relate to the main dynamical components of the pathway. These are illustrated in the example of Figure 1 and include arrival rates, service point active LOS and capacity, and discharge routing probabilities. Not all service points on a patient pathway are appropriate to be modelled. This could be because they do not contain blocking after service, and are therefore out of scope of this paper (e.g. above example of elective pathway), or because delays are not based on capacity per se (e.g. social care package dependence on home modifications), or because practically there is a lack of data or understanding of their function (e.g. a generalised service which accepts patients from a variety of pathways). The model should therefore concern a *modelled pathway* accounting for appropriate service points where capacity-related dynamics are explicitly modelled. Furthermore the model should make suitable assumptions for interactions between service points on and immediately off this modelled pathway, i.e. it is not necessarily realistic to assume a zero delay to a service point off the modelled pathway. Instead, some empirically-derived (capacity-independent) delay could be assumed, such as for social care as illustrated in Figure 1 (note this idea is suggested, but not developed, in Monks et al, 2015 when studying a similar problem).

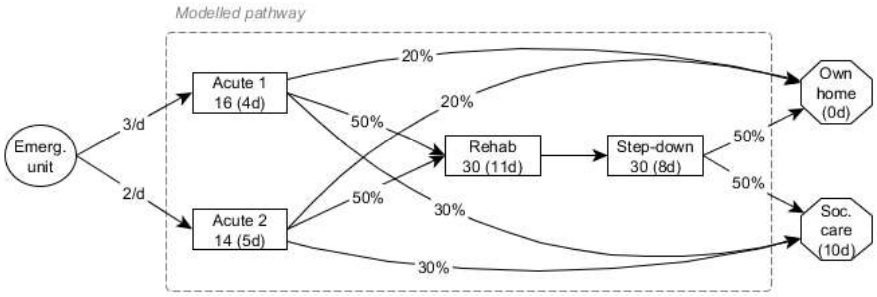


Figure 1 Illustration of an example patient pathway, detailing daily arrival rates from an emergency unit; discharge routing probabilities as percentages; capacity and average active length of stay (in parentheses) for service points on the modelled pathway; and average delays to discharge for transfers to service points off the modelled pathway (in parentheses)

Above all, the model should be able to represent the current pathway, whilst also being able to evaluate hypothetical future pathways through *what if* scenario analysis. Strategic changes to the structure of the pathway could include new service points being added or discharge routing probabilities being modified to take account of, e.g., greater discharges to social care due to enhanced responsibilities. Scenarios to reflect more operational concerns could simply involve increasing or decreasing the capacity of current service points in targeting some desired performance measures, e.g. that no more than 1 in N patients have to wait on arrival. Scenarios could

also directly target blocked LOS where capacity is optimised in order to reduce the average proportion of blocked service channels at a service point by x%.

The model must be readily useable by health analysts and service managers, and thus must satisfy certain criteria which separate it from a purely academic exercise. It must be sufficiently simple and intuitive to all those involved in using it, especially with regards to inputs and outputs. The former must be clearly associated with real-life components and on-the-ground levers that are familiar to users and for which data is available. The latter must be composed in a way as to appreciate the realistic stochastic nature of the pathway dynamics and contain, for instance, quantiles of the distribution of service point waiting list size, such that users can identify where bottlenecks would form. The outputs must also enable evaluation of cost-efficacy and be versatile to users' demands for what measures are included. The model itself should be sufficiently flexible to account for the variability of parameters over time which can be realistically expected, e.g. increasing and seasonal arrival rates and discharge restrictions at weekends. Finally, the solution must not involve the use of costly software: promoting take-up of the model relies upon as few barriers as possible for practitioners to evaluate it and ultimately use it (Fletcher & Worthington, 2009).

1.2 Current state of the art

It has long been the case that NHS capacity planning has relied on deterministic, or *averages-based*, methods (Harper, 2002) and in the authors' and other's (Monks et al, 2016) experiences not a lot has changed to date. Such methods are reasonable if there is no variability in the pathway, i.e. if there are constant arrival rates and lengths of stay. However, these assumptions are rarely valid and when violated the resulting capacity requirements are under-estimated, sometimes by up to 40% (Monks et al, 2016). Furthermore, without a stochastic model the mechanics of blocking after service cannot be appreciated, and thus cannot allow capacity planners to evaluate the effect of capacity-related interventions on blocked LOS.

In the published literature many investigators use discrete event simulation or analytical queuing theoretic methods to consider the detailed capacity dynamics of service points from intensive care (Griffiths et al, 2006), to emergency units (Mayhew & Smith, 2008) and rehabilitation wards (Wood et al, 2014a). Of those that look at a pathway over multiple service points, many make use of system dynamics, such as Brailsford et al, 2004 and Rashwan et al, 2015 who investigate patient flow around a health and social care system. However, in using system dynamics, or *continuous simulation*, the pathway is not considered stochastically and is thus at risk of under-estimating required capacity.

There are a limited number of published studies which both model patient flow around a network of service points *and* do this stochastically. This is typically performed through discrete event simulation using commercial software such as Simul8, Arena, and Anylogic (Pitt, 1997, El-Darzi et al, 1998, Cahill & Render, 1999, Moreno et al, 1999, Blasak et al, 2003, Bayer et al, 2010, Cordeaux et al, 2011, Brailsford et al, 2013, Monks et al, 2016, Rodrigues et al, 2017). For reasons of tractability, analytical models have tended to be restricted to smaller and bespoke networks of two or three service points (Harrison, 2001, De Bruin et al, 2007, Lin et al, 2014), with little to no flexibility to expand the network considered. Typically, blocking after service has not been appreciated in models (Monks et al, 2016), despite being an inherent and recognised (Monks et al, 2015) aspect of the pathway investigated. There are notable exceptions, however, including efforts to assess the effect of blocking after service in networks of three or more service points by both simulation and analytical methods (Koizumi et al, 2005, Cochran & Bharti, 2006, Osorio & Bierlaire, 2009, Bretthauer et al, 2011, Griffin et al, 2012). However, in each case these either require the use of commercial software or demand high levels of mathematical expertise by users. There have been a number of attempts to develop and distribute multi-use modelling packages, but in each case they appear to have required some model-building on the part of the user and evidence of their continued availability could not be found (Harper, 2002, Vasilakis et al, 2013).

Included in Appendix A is an itemised review of these papers against criteria based on the modelling requirements outlined in Section 1.1. It can be seen that there is no one paper published to date which satisfies the objectives of providing a reliable and versatile capacity model for a patient pathway, which does not make use of costly and technical software and which is not too complicated for health analysts (whose quantitative analytical skillset can be lacking – Bardsley, 2016).

In this paper we address this gap in the literature; developing an easy-to-use tool coded in the free software "R" which offers an advancement over the averages-based approaches commonly used on the ground. The remainder of this paper is structured as follows. Section 2 specifies the problem mathematically and details the solution approach taken. In Section 3 the ease of practical application of the model is demonstrated through an

illustrative example concerning the redesign of a stroke service. And finally a discussion of results, limitations, and extensibility is provided in Section 4.

2. Approach

2.1 Framing the problem

Queuing systems can be mathematically framed in a number of ways (see Gross et al, 2008 for a thorough account of queuing theoretic problems and solution methods). Of these, a typical choice is through Markov modelling, due to a natural fit to the problem as is now explained through a simple example.

A Markov chain is a stochastic process that satisfies the memorylessness property, i.e. that movement to future states of the chain is dependent only on the current state and not of those visited previously. In a continuous-time Markov chain the holding times within each state are therefore exponentially distributed, since this is the only continuous-valued distribution for which the memorylessness property holds. A simple example of a continuous-time Markov chain is the birth-death process, where transitions between states of the chain are based on either a birth, with rate λ, or death, with rate μ. This process is depicted in Figure 2.

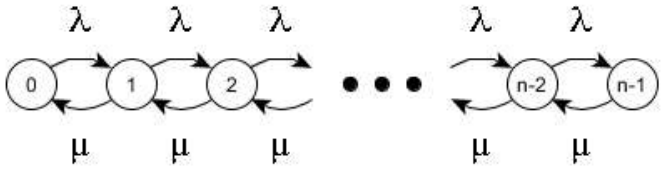


Figure 2 A birth-death process with birth rate λ and death rate μ

The process described in Figure 2 is a finite-state non-absorbing Markov chain. It is finite-state because there are n distinct states of the chain, whose state space is thus defined as $S = \{0,1,\dots,n-1\}$. It is non-absorbing since there is no one state that can be reached from which movement out is impossible – that is, the process will continue indefinitely. The $n \times n$ transition rate matrix of this system is written as

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ \mu & -(\lambda+\mu) & \lambda & & 0 \\ 0 & \mu & -(\lambda+\mu) & & \vdots \\ \vdots & & & \ddots & \lambda \\ 0 & 0 & \dots & \mu & -\mu \end{bmatrix}.$$

This simple Markov process can be used to model a variety of real-life phenomena, such as the growth of bacteria, or indeed the number of patients in the rather trivial example of a single bed hospital unit, where λ and μ^{-1} represent means of the exponentially-distributed arrival rate and LOS respectively and the number waiting is given by $max(0, s-1)$ where $s \in S$ is the state of the system. Note the randomness conferred through an exponential arrival rate is a quite fitting property since realistically there is typically no dependence of one arrival from another. For LOS though, it is not always an appropriate assumption, as is discussed later.

Long-run, or steady state, probabilities of the process occupying each state $s \in S$ can be deduced through the solution of $\boldsymbol{\pi} \cdot Q = \mathbf{0}$ where $\boldsymbol{\pi} = [\pi_0 \quad \pi_1 \quad \cdots \quad \pi_{n-1}]$ are the state probabilities and $\sum_{s \in S} \pi_s = 1$. Thus the probability of an arriving patient finding the hospital empty is π_0 and the probability of having to wait is $1 - \pi_0$. Other performance measures such as waiting times can be obtained through the steady-state probabilities using standard results such as Little's Law.

In this study the admitted patient pathway is modelled as a continuous-time Markov chain, under exactly the same principles as this simple example, albeit with a more practicable approach to capture the additional complexities. But first, it's necessary to frame the state space of the network and describe the possible state transitions.

Let $F = \{f_1, f_2, \dots, f_N\}$ be the set of all N service points on the modelled pathway, where each $f \in F$ has c_f service channels. The number of service channels – or *capacity* – is defined by the number of patients which can concurrently be served at the service point (for inpatient service points this is simply the number of beds). The

distribution of service channel active LOS within service points must approximate well to the respective empirical distributions and, as mentioned earlier, the exponential distribution is not usually appropriate to this end. Although with an extra parameter a good fit can usually be obtained with two-parameter distributions such as the log-normal or gamma (Marazzi et al, 1998). However, these are unsuitable candidates for a Markov model since they fail to satisfy the memorylessness property. Instead, active LOS is here modelled through a distribution which is based on exponential "building blocks".

The Erlang (k, μ) distribution is a special case of the gamma distribution where the shape parameter, k, is restricted to positive integers. It can be fitted to data through the standard techniques, such as maximising the likelihood function or matching theoretical and empirical moments or quantiles. Previous studies highlight its applicability to hospital LOS (Kolker, 2008). Conveniently, the distribution can itself be interpreted as an acyclic absorbing Markov process; a series of k exponential distributions, or *phases*, each with rate parameter μ (see Figure 3). Thus, use of this distribution can not only provide a good fit to active LOS data but, crucially, permits incorporation within the wider non-absorbing Markov chain used to model the pathway.

The Erlang distribution is also used to model the delay in discharge to service points off the modelled pathway whose capacity is not considered but for which delays can realistically exist (appreciating such delays through a random variable is suggested, but not developed, in Monks et al, 2015). The patient is thus held at the service point, occupying a service channel, until all exponential phases have been transitioned through. Let $\bar{F} = \{\bar{f}_1, \bar{f}_2, \dots, \bar{f}_{\bar{N}}\}$ be the set of all M discharge destinations off the modelled pathway, and such that $F \cap \bar{F} = \emptyset$.

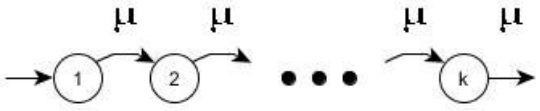


Figure 3 The Erlang (k, μ) distribution with k phases each with rate parameter μ

The state representation for the system, S, is a combination of the state representations, s_f, for each service point $f \in F$ such that $S = \{s_{f_1}, s_{f_2}, \dots, s_{f_N}\}$. The minimal representation of each constituent s_f is given by the components detailed in Table 1. Thus, the minimal state representation necessary to describe the system for each service point $f \in F$ is $s_f = \left\{\xi_f, \eta_{f,a_f}, \omega_{f,g_f}, \psi_{f,\bar{g}_f,b_f}\right\} \quad \forall a, g, \bar{g}, b$.

Table 1 State representation of the system

Component	Description
ξ_f	The number waiting externally for service point $f \in F$, i.e. the number waiting who are not already at other (upstream) service points on the pathway
η_{f,a_f}	The number of patients at service point $f \in F$ in exponential phase $a_f \in \{1,2,\dots,k_f\}$ of their active LOS, where k_f is the number of phases of the Erlang-distributed active LOS at service point f
ω_{f,g_f}	The number of patients at service point $f \in F$ waiting for discharge to service point $g_f \in G_f = \{g_{f,1}, \dots, g_{f,M_f}\}$, where $G_f \subset F$ are the pathway downstream destinations accessible from service point f, and M_f is the total number of such destinations
$\psi_{f,\bar{g}_f,b_f}$	The number of patients at service point $f \in F$ in phase $b_f \in \{1,2,\dots,p_{\bar{g}_f}\}$ of waiting for discharge off the pathway to external destination $\bar{g}_f \in \bar{G}_f = \{\bar{g}_{f,1}, \dots, \bar{g}_{f,\bar{M}_f}\}$, where $\bar{G}_f \subset \bar{F}$ are the external destinations accessible from service point f, $\bar{M}_f$ are the total number of such destinations, and $p_{\bar{g}_f}$ are the numbers of phases of the Erlang-distributed delay distributions to external destination $\bar{g}_f$

Movement around this state space is analogous to discrete transitions in the continuous-time Markov chain. Each transition is tied to one of three types of event: external arrivals, active LOS phase completions, and external delay phase completions. At any one of these events there may be a number of connected patient movements. For example, take an active LOS phase completion. If the phase completed is not the final phase (i.e. $a < k$) then there are no associated patient movements. However, if the phase completed is the final phase of active LOS (i.e. $a = k$) and there is a vacant service channel at the probabilistically-selected discharge destination, then the patient is discharged and a waiting patient admitted to the now-vacated service channel

(provided one is waiting). This may have a further knock-on effect as the admitted patient frees up a service channel upstream, and so on. Conversely, if a patient completes their final active LOS phase and there is no downstream capacity then they remain in their service channel in a holding stage (i.e. ω) until the selected downstream service point has availability through a patient discharge. For more information on the range of possible patient movements connected to each type of state transition see Appendix B.

At each state transition the type of event which occurs depends on the occupied state (defined through Table 1) as well as some rates associated with the three types of event. For example, the relative likelihood of a state transition involving an active LOS service phase completion in a particular service point is doubled should there be two patients rather than one occupying that phase of service. However, the absolute likelihood of that event occurring is of course dependent on the rates of the other events. Considering a trivial pathway with only one service point, external arrival rate λ, and two service channels with a single active LOS phase of rate μ, the probability of a state transition involving a service completion would be $2\mu/(2\mu + \lambda)$ if both channels are occupied, $\mu/(\mu + \lambda)$ if one is occupied and zero if none are occupied. For the modelled pathway considered here, the rates and probabilities that, together with Table 1, complete the specification of the system are detailed in Table 2. Note that, arrivals rates are independent of occupied state, unlike the service and delay phase rates.

Table 2 System rates and probabilities

Component	Description
λ_f	External arrival rate into service point $f \in F$
μ_f	Phase completion rate for each of k_f active LOS phases for service point $f \in F$
Φ_f	Vector of probabilities of discharge to destinations G_f and $\bar{G}_f$ such that $\Phi_f = \left\{\phi_{G_f}, \phi_{\bar{G}_f}\right\} = \left\{\phi_{g_1}, \dots, \phi_{g_{M_f}}, \phi_{\bar{g}_1}, \dots, \phi_{\bar{g}_{\bar{M}_f}}\right\}$ and $\sum \Phi_f = 1$
$\gamma_{\bar{f}}$	Phase completion rate for each of $p_{\bar{f}}$ delay phases to external destination $\bar{f} \in \bar{F}$

2.2 Solving the problem

The stochastic solution to a queuing system can be approached analytically or through simulation. Analytical methods involve setting up a mathematical description of the system which can then be solved either exactly or via numerical approximation, depending on complexity. Analytical solutions to queueing systems may make use of Laplace transforms, such as through the Pollaczek-Khintchine formula (Vidyarthi & Kuzgunkaya, 2015), or matrix analytic methods, in solving systems based on Markov-chains (e.g. Griffiths et al, 2013). Analytical approximations can also be attempted, such as in Bretthauer et al, 2011 when modelling a patient pathway, albeit typically with limitations or range restrictions on parameter values. Simulation is an alternative solution approach based on computationally performing a number of runs which each details a possible set of operational events which could unfold over a specified time period. Simulation has been gaining much traction in healthcare in recent years, due to advances in computational power and software available (Gunal & Pidd, 2010).

In this study simulation is chosen over an analytical solution due to flexibility and tractability. It is useful to have flexibility should, for example, arrival rates be required to change over time to appreciate seasonality (transient solutions, required for time-inhomogeneous Markov chains, are notoriously more complex than steady-state results). But the principal reason is tractability. An analytical solution based on matrix methods would require the manipulation of a very large and irreducible square transition rate matrix. Let there be x patients in service point f, where $x \leq c_f$. For each individual, they may reside in either one of the k_f active LOS phases, or one of M_f queues for downstream pathway service points, or one of $\sum_{j=1}^{\bar{M}_f} p_{\bar{g}_{f,j}}$ delay phases for external discharge destinations. Thus, using multiset notation there are $\left(\binom{x+1}{y_f-1}\right)$ unique positions for each patient where $y_f = k_f + M_f + \sum_{j=1}^{\bar{M}_f} p_{\bar{g}_{f,j}}$. It follows that the total number of states required to define each unique representation of the system is therefore $\prod_{f\in F}\sum_{x=0}^{c_f}\left(\binom{x+1}{y_f-1}\right)$. So for the example pathway of Figure 1, and assuming just one active LOS phase for each service point on the modelled pathway and one delay phase to social care, the dimension of the matrix to be manipulated would be of 11 orders of magnitude (noting this is in fact a lower bound since the state space would also require accounting for patients in the external queues for the two acute service points).

The approach taken here, based on Wood et al, 2014b, works by simulating movements around the state space of the Markov chain by dynamically sampling the next state as one accessible from the neighbourhood of the current. At each such iteration there is a two-step process: the first step determines the event and the time until the event, and the second step updates the state vector S to take account of the event occurring (noting that due to the memorylessness property it is, of course, not necessary to track times in each state). With a specification of the accessible, or neighbouring, states deducible from the current state vector, S, the rates of each event can be determined. For example, the rate of a first phase active LOS completion in service point f is $\eta_{f,1} \cdot \mu_f$, which is the product of the number in that phase and the individual phase rate (as in Table 2). These rates are used to draw random samples from the exponential distribution for each event, e.g. $-(\eta_{f,1} \cdot \mu_f)^{-1} \log(1-u)$ where u is a random number in the $U(0,1)$ uniform distribution. These give the unconditional simulated times until the occurrence of each event, from which the event associated with the smallest time is selected (first step). All that is required then is to update the state vector S to take account of this state transition (second step). If, in the above example of the first phase active LOS completion, this event has the lowest simulated time then the new state vector S would be one with $\eta_{f,1}$ reduced by one and $\eta_{f,2}$ increased by one (assuming $k_f > 1$).

To begin, a starting point in the state space must be selected through specifying the state vector S. The simplest option is to assume an empty system, in which case the value of all elements of S is zero (alternatively the η can be pre-populated corresponding to some current occupancy figures). From this point the two-step process mentioned above is performed until some terminating condition is met, such as the run time reaching one year (following a suitable warm-up period). This is one run of the simulation and describes just one possible way events could unfold over the course of the simulated period. In order to capture the many other possibilities a number of runs are performed. If R runs are performed where for each run $r \in \{1, \dots, R\}$ there are n_r events, then the ultimate raw model output is a list of R matrices, each with n_r rows and $|S| = N + \sum_{f \in F} y_f$ columns (recalling $N = |F|$). On each matrix an additional column is used to keep track of the time at which each event has occurred. From this can be calculated the total times for which each unique state of the state space has been occupied which, in turn, yields performance measures and costs that can be aggregated (averaged) over all R runs. This is demonstrated in the next section through a real-life example.

3. Application and results

Development of the model has been prompted by the re-design of a stroke pathway being considered in a region of England containing approximately one million people. Currently stroke services are provided in three district general hospitals, but there is a desire to investigate whether centralising could bring the benefits supported by recent studies promoting *hyper-acute* services (Liu et al, 2011). To help assess this, the project would like to know how much capacity would likely be required along such a hypothetical *future state* pathway.

The model is written in the freely-available open-source statistical software "R" and has been cross-verified in the commercial Anylogic program. The script and input files used in the example described forthwith are available in the Supplementary Material.

3.1 Current state

Before any scenario analysis can be performed, it is first necessary to model the *current state* pathway to get a baseline. Service points on the modelled pathway include two acute units and two rehab units at hospitals X and Y and one combined unit at hospital Z. Movement off the modelled pathway is either through death or discharge to the patient's own home or a bedded nursing/residential home. A delay to discharge is typically experienced if social care support is required at the patient's own home or if they are transferred to a bedded care facility. These discharge destinations are not on the modelled pathway, firstly, because of incomplete data and understanding of their dynamics (due in part to their being a generalised service accepting patients from all kinds of pathways), and secondly, because delays to them are not always capacity related (these can be due to patient/family deliberation over choice of nursing/residential home or due to equipment installation in the patient's home as part of a social care package). It is, however, necessary to take account of these delays – which can be significant – and in doing so makes use of an important feature of this model where non capacity-related blocking after service can be appreciated. Through available data collected for the project the current state pathway can be mapped (Figure 4A), and model inputs determined (Table 3A). These inputs are stored as a simple *csv* file to be read in via the model script (see Supplementary Material).

Table 3A has $|F| + |\bar{F}| = 5 + 5 = 10$ rows with the left-most 10×10 square containing the transition matrix elements Φ_f and subsequent columns containing the daily arrival rates λ_f, the number of service channels c_f,

and the mean active LOSs k_f/μ_f (i.e. the Erlang mean), in days, for the k_f respective phases. Following are the mean delays to discharge $p_{\bar{f}}/\gamma_{\bar{f}}$, in days, for the $p_{\bar{f}}$ delay phases. In fitting the distributions for active LOS and delay to discharge the Akaike information criterion is used to ensure that any inclusion of an additional parameter associated with an Erlang distribution with a shape parameter greater than one is warranted. See Appendix C for more detail on this fitting process.

The modelled outputs, produced from 10,000 runs each with a warm-up period of 100 days, are summarised in Table 4A (note these are written from the script alongside more-detailed output data and plots for each individual run). The results support validity of the model in appreciating the most recognised aspects of current system behaviour. This includes substantial levels of blocking at hospital Z where, on average, one in every four beds occupied are blocked (rising to at least one in every two on one in 20 days), and high utilisation of the rehabilitation unit at hospital X where, on average, four in every five beds are occupied and one in every four referrals have to wait, causing upstream blockages in the acute unit.

3.2 Future state

Attention is now turned to examining the alternate pathway configuration. This involves transforming acute provision by way of centralisation to one site and through addition of a short-stay hyper-acute stroke unit (HASU) acting as a singular point of access to stroke-specific care.

Parameters associated with this pathway have been estimated by the clinically-led project working group. These have been grounded on the current state and influenced by the experiences of other health systems which have moved to a HASU setup, such as London and Manchester where there is evidence of reduced mortality and LOS (Morris et al, 2014) and improved interventions (Ramsay et al, 2015) leading to better outcomes. The only parameters not included in the pathway map (Figure 4B) and model inputs (Table 3B) are the numbers of beds (service channels), since these are to be targeted in arriving at a system that is appropriately resourced.

This requires an assessment of the performance measures associated with a number of scenarios for which bed numbers are specified. To exemplify the perils of deterministic capacity modelling, the averages-based bed numbers are used as a starting point. Calculating capacity as the product of the average arrival rate and average LOS yields an estimated requirement for 11, 12, 7 and 4 beds at the HASU, acute stroke unit (ASU) and rehabilitation units at X and Y respectively (e.g. $\lceil 4.1 \times 2.5 \rceil = 11$ for the HASU). Modelled results describe an unstable queuing system in which the front door queue for the HASU grows ever larger with time – thus confirming the finding of Monks et al, 2016 that averages-based methods "*substantially under-estimate capacity requirements*".

Capacity can thereafter be increased from this starting point until some performance criteria are met, within any operational or cost constraints. This can be framed as a multi-objective optimisation problem, where some trade-off is sought between the improving performance and rising costs associated with increasing capacity (perhaps against intuition, reducing blocking to zero is unlikely financially optimal). Here, however, capacity is simply targeted at the minimum required to achieve some desired performance measures; namely that no more than 1 in 100 days should a HASU arrival have to wait, no more than 1 in 20 should an ASU arrival wait, and no more than 1 in 10 should an arrival at either rehabilitation unit have to wait (i.e. 90% of the time there should be at least one bed available). The bed numbers that achieve this are found to be 20 for the HASU, 23 for the ASU, and 13 and 9 for rehabilitation units X and Y respectively – see Table 4B for the associated performance measures.

Note that the probability of non-zero wait is quite a bit below the 1 in 10 day threshold for the rehabilitation units. If, however, the bed numbers are incrementally reduced in either or both of these then the threshold would be violated for the ASU. This is because of the increased utilisation of the rehabilitation units pushing up the probability of delay to admission from the upstream ASU, consequently increasing ASU utilisation (due to patients staying longer) and thus too the probability of an arrival finding the ASU full and having to wait. This is a clear illustration of why it is vitally important to model separately active and blocked LOS, with the latter dependent on the dynamics of the system. Also demonstrated is the importance of modelling variability: averages-based capacity estimations would have under-estimated total capacity by 48% (i.e. 34 *c.f.* 65 beds).

It is also worth comparing capacity between the current and future states, where the modelled estimates suggest a two-fifth reduced bed requirement. But while there is of course capacity relief in centralising from queuing "economies of scale", there are other factors which detract from a like-for-like comparison. In augmenting the pathway for the HASU system, assumptions have been made regarding patient flows between service points

both on and off the modelled pathway, where a key assumption is that (39%) more patients are discharged to an improved social care service. So where capacity is reduced on the modelled pathway, consideration should be given to whether it needs to be increased off it in order to achieve the specified delay distributions to the out-of-hospital providers.

Any further increases to capacity for the out-of-hospital providers off the modelled pathway could bring about reductions in capacity requirements on the modelled pathway. If delays to discharge to social care and nursing/residential homes are halved (to averages of 1.5 and 8d respectively) then similar performance measures can be attained with 20, 21, 12 and 8 beds at the HASU, ASU, and X and Y rehabilitation units respectively. That is four fewer, or 6% less, beds required. Delays to discharge could be reduced not only through increasing downstream capacity however, as there could be other reasons for delay. For example, patients can experience discharge delays to social care whilst mobility equipment is being installed at their home. Such delays can be mitigated if reliable estimates of discharge readiness can be communicated early on (active LOS estimates can be obtained using regression methods, as in Franchignoni et al, 1998). Reducing such delays through more investment in social care and better co-operation with acute units are a key focus in the UK health service (Edwards, 2014).

Table 3 Model input for Current (A) and Future State pathway (B)

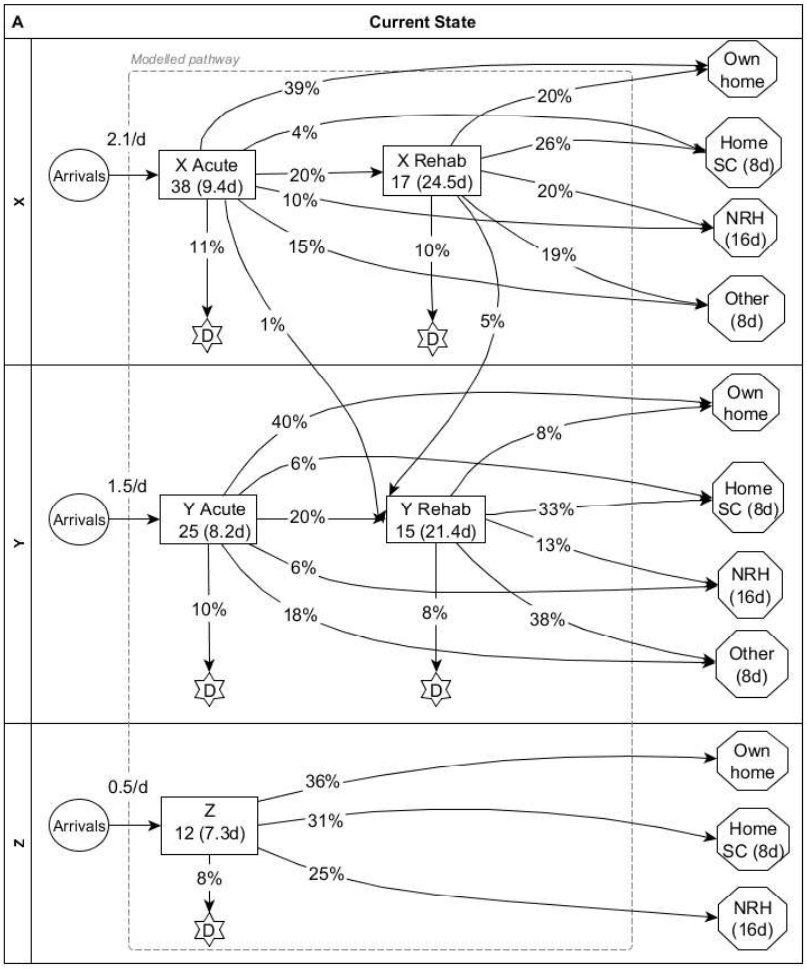
A	AC X	AC Y	Z	RH X	RH Y	Death	Home	Home SC	N/R H	Other	λ	c	k/μ	k	p/γ	p
AC X				0.20	0.01	0.11	0.39	0.04	0.10	0.15	2.1	38	9.4	1		
AC Y					0.20	0.10	0.40	0.06	0.06	0.18	1.5	25	8.2	1		
Z						0.08	0.36	0.31	0.25		0.5	12	7.3	1		
RH X					0.05	0.10	0.20	0.26	0.20	0.19		17	24.5	1		
RH Y						0.08	0.08	0.33	0.13	0.38		15	21.4	1		
Death																
Home																
Home SC															8	1
N/R H															16	1
Other															8	1

B	HASU	ASU	RH X	RH Y	Death	Home	Home SC	N/R H	λ	c	k/μ	k	p/γ	p
HASU		0.40			0.08	0.52			4.1	?	2.5	1		
ASU			0.15	0.10	0.04	0.30	0.33	0.08		?	6.8	1		
RH X					0.10	0.10	0.50	0.3		?	24.5	1		
RH Y					0.08	0.10	0.50	0.32		?	21.4	1		
Death														
Home														
Home SC													3	1
N/R H													16	1

Table 4 Modelled outputs of Current (A) and Future State pathway (B)

A	Probability of non-zero wait	Occupied service channels				% occupied beds blocked	
		Mean	85%	95%	99%	Mean	95%
AC X	0.05	27.0	32.6	35.5	37.3	19%	33%
AC Y	0.04	16.8	21.1	23.5	24.7	18%	33%
Z	0.06	6.9	9.7	11.1	11.8	27%	52%
RH X	0.25	13.3	16.4	16.9	17.0	17%	34%
RH Y	0.10	10.0	13.2	14.4	14.8	18%	36%

B	Probability of non-zero wait	Occupied service channels				% occupied beds blocked	
		Mean	85%	95%	99%	Mean	95%
HASU	0.01	10.3	13.6	15.8	18.4	0%	1%
ASU	0.04	14.9	19.0	21.3	22.7	17%	31%
RH X	0.05	7.6	10.4	11.7	12.4	12%	29%
RH Y	0.05	4.6	6.7	7.8	8.5	12%	32%



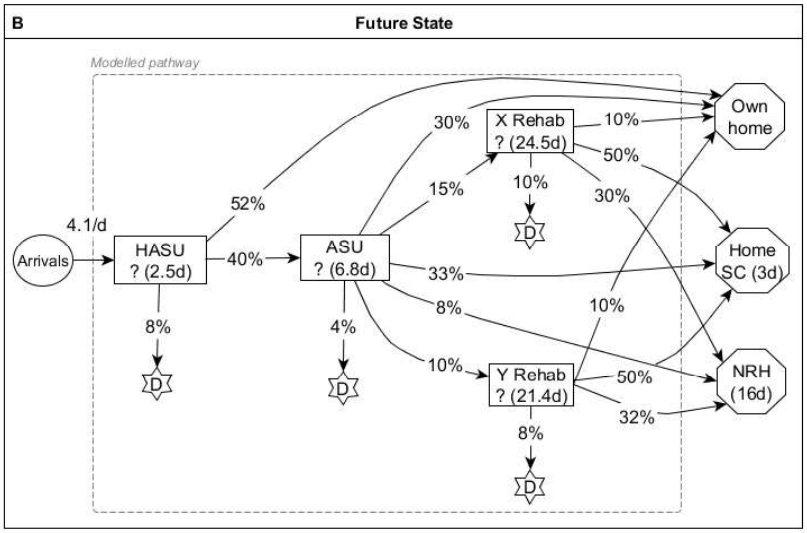


Figure 4 Capacity-flow maps of Current (A) and Future State pathway (B), detailing daily arrival rates onto the modelled pathways; discharge routing probabilities as percentages; capacity and average active length of stay (in parentheses) for service points on the modelled pathway; and average delays to discharge for transfers to service points off the modelled pathway (in parentheses)

4. Discussion

This paper builds upon established queuing theoretic principles in addressing a key gap in the literature concerning the capacity-related dynamics along a patient pathway with delays to transfer and discharge. A conceptually appropriate theoretical model and a practically suitable solution are presented, which are thereafter applied to a stroke service evaluation. Attention is now turned to a discussion on the merits and limitations of this approach and opportunities for future use and development.

It is perhaps first worth commenting on some lessons learned from providing modelling support to this project. To begin with it is important to clearly understand what should and should not be under the modelling remit. To ensure the model is useful it must be accurate, and the focus in model selection should be on ensuring any candidate solutions fit the problem and not vice versa. Data and results should be shared with clinicians and service managers as this can act as crucial validation. For instance, relaying summaries of the data used for calibration can validate input parameters, and seeking recognition of modelled outputs (such as mean and 90% occupancy rates) can confirm credibility of the model construct in approximating real life behaviour. Lack of such communication between modellers and stakeholders has been recognised as the "*primary factor contributing the most to poor stakeholder engagement in healthcare simulation projects*" (Jahangirian et al, 2015).

Turning to pros, a critical asset of the model is that practically it is actually quite straightforward and can be readily understood and operated by healthcare practitioners. This is because the model has been clinically-shaped from the outset: it has been specifically designed to incorporate variables felt to be intuitively relevant and to output the relevant metrics sought for inspection. But whilst appearing on the surface simple and intuitive the mathematical underpinnings are sufficiently sophisticated in equating to the quite complex dynamics at play. This is not always appreciated in modelled solutions. In also investigating a stroke pathway, Monks et al, 2016 make the simplifying assumption that capacity is infinite and so they "*cannot predict the length of a delay a patient experiences*". Yet the crux of modelling an admitted patient pathway is precisely these delays and the inefficiencies they cause through the blocking of service channels for others (as mentioned in Section 1 – a very substantial problem in the UK NHS currently). The model presented here considers the delayed component of LOS as a dependent variable on downstream capacity and so is able to meaningfully evaluate what if scenarios without having to make the unrealistic assumption that total LOS is unaffected under any hypothetical scenario considered. Furthermore, delays off the modelled pathway are appreciated in line with suggestions from Monks et al, 2015 that "*instead of modelling social services in detail, such delays in discharge from rehabilitation could be incorporated as a random variable*". These advantages help set apart the model presented here from other efforts where Brailsford et al, 2009 find that only 1 in 20 simulation models in health are practically implemented.

While there are benefits in having implemented the model in open-source software – in respect of cost, training, and control over outputs – there are also drawbacks. Some of the commercially available programs have impressive ability to graphically display simulated movements around the system over time. This is a most useful function for quickly attracting interest and succinctly explaining the principles of simulation to those unfamiliar. However, this is not a necessary feature for longer-term working with a project group beyond the first meeting where simulation can be more thoroughly introduced. Commercial software also benefits from versatility in the range of queuing networks that can be constructed, e.g. combining admitted and non-admitted pathways. Whilst it would be possible to generalise further the model presented here to cater for such versatility it would require more inputs than currently captured (Table 3), which would detract from the afore-mentioned benefits of user simplicity. It should be noted that the software used to code the model – "R" – is gaining traction in the NHS, where there is a national initiative to train analysts in its use (The Health Foundation, 2018).

Possible limitations in light of modelled assumptions are now discussed. One such assumption is that transfers occur instantaneously: as soon as a patient is discharged then a waiting patient is admitted. If the two service points are not close-by and travel times are significant then this could potentially underestimate required capacity. Another assumption is that arrivals are time-homogenous, when in reality there may be seasonality with respect to month in year, day in week, and hour in day. However, due to the versatile nature of the dynamic neighbourhood simulation approach, the code can readily be adjusted to take account of this by simply conditioning the value of λ_f on the simulated time at that iteration. This can also be used to incorporate an increasing arrival rate, for instance, to take account of an ageing/growing population. Active LOS is assumed Erlang-distributed in the model which, whilst affording a fair degree of flexibility (Appendix C), is not as flexible as some of the more generalised *phase-type* distributions which have been shown to approximate well to stroke data (Vasilakis & Marshall, 2005). Owing to the underlying Markov framework of the model presented here, extension to these distributions could be conveniently facilitated in future work (however, this would be at the expense of user take-up due to additional complexity in required parameters). Another assumption made is that that once the queue is entered it cannot be left other than through admission. Should this assumption not hold then further work could be conducted to ensure *reneging* is appreciated. The related queuing phenomena of *balking* is incorporated within the model code to the extent of being able to toggle between allowing external arrivals to wait or alternatively configuring a loss system, where external arrivals are "lost" if there is no available service channel (noting this latter option is not appropriate to the stroke project considered here).

In summary, this paper presents a versatile capacity model for an admitted patient pathway, bridging a gap in the literature and providing a readily useable solution to healthcare professionals, as demonstrated here in a stroke setting. The model is also in the process of being calibrated for geriatric community capacity estimation, and is intended to be distributed to other healthcare organisations as a means to promote the afore-mentioned benefits of system-wide thinking. Through greater appreciation of the per-bed cost and resource requirements the model can provide a framework for detailed understanding of the balance between higher costs and improved fluidity associated with increased capacity along a pathway, which could be used to address current thinking on the importance of community capacity in alleviating blockages at more expensive upstream points.

References

Bardsley, M. (2016). Understanding analytical capability in health care. Do we have more data than insight. London: The Health Foundation.

Bayer, S., Petsoulas, C., Cox, B., Honeyman, A., Barlow, J. (2010). Facilitating stroke care planning through simulation modelling. *Health informatics journal*. 16 (2).

Blasak, R. E., Starks, D. W., Armel, W. S., Hayduk, M. C. (2003). Healthcare process analysis: the use of simulation to evaluate hospital operations between the emergency department and a medical telemetry unit. *Proceedings of the 35th conference on Winter simulation*. Winter Simulation Conference.

Brailsford, S.C., Lattimer, V.A., Tarnaras, P., Turnbull, J.C. (2004). Emergency and on-demand health care: modelling a large complex system. *Journal of the Operational Research Society*. 55 (1).

Brailsford, S. C., Harper, P. R., Patel, B., Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*. 3 (3).

Brailsford, S.C., Desai, S.M., Viana, J. (2010). Towards the holy grail: Combining system dynamics and discrete-event simulation in healthcare. *Proceedings of the 2010 Winter Simulation Conference*.

Brailsford, S. C., Bolt, T. B., Bucci, G., Chaussalet, T. M., Connell, N. A., Harper, P. R., Klein, J.H., Taylor, M. (2013). Overcoming the barriers: a qualitative study of simulation adoption in the NHS. *Journal of the Operational Research Society*. 64 (2).

Bretthauer, K. M., Heese, H. S., Pun, H., Coe, E. (2011). Blocking in healthcare operations: a new heuristic and an application. *Production and Operations Management*. 20 (3).

Cahill, W., Render, M. (1999). Dynamic simulation modeling of ICU bed availability. Proceedings of the 31st conference on Winter simulation.

Cochran, J. K., Bharti, A. (2006). Stochastic bed balancing of an obstetrics hospital. *Health care management science*. 9 (1).

Cordeaux, C., Hughes, A., Elder, M. (2011). Simulating the impact of change: implementing best practice in stroke care. *London journal of primary care*. 4 (1).

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 (2).

Edwards, N. (2014). Community services: How they can transform care. *The King's Fund*.

El-Darzi, E., Vasilakis, C., Chaussalet, T., Millard, P. H. (1998). A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health care management science*. 1 (2).

Fletcher, A., & Worthington, D. (2009). What is a 'generic' hospital model? – a comparison of 'generic' and 'specific' hospital models of emergency patient flows. *Health Care Management Science*. 12 (4).

Franchignoni, F., Tesio, L., Martino, M.T., Benevolo, E., Castagna, M. (1998). Length of stay of stroke rehabilitation inpatients: prediction through the functional independence measure. *Annali dell'Istituto Superiore Di Sanita*. 34 (4).

Griffin, J., Xia, S., Peng, S., Keskinocak, P. (2012). Improving patient flow in an obstetric unit. *Health care management science*. 15 (1).

Griffiths, J.D., Price-Lloyd, N., Smithies, M., Williams, J. (2006). A queueing model of activities in an intensive care unit. *IMA Journal of Management Mathematics*. 17 (3).

Griffiths, J.D., Williams, J.E., Wood, R.M. (2013). Modelling activities at a neurological rehabilitation unit. *European Journal of Operational Research*. 226 (2).

Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M. (2008). Fundamentals of queuing theory. Wiley Interscience.

Gunal, M.M., Pidd, M. (2010). Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*. 4 (1).

Harper, P. R. (2002). A framework for operational modelling of hospital resources. *Health care management science*. 5 (3).

Harrison, G. W. (2001). Implications of mixed exponential occupancy distributions and patient flow models for health care planning. *Health Care Management Science*. 4 (1).

Jahangirian, M., Taylor, S. J., Eatock, J., Stergioulas, L. K., Taylor, P. M. (2015). Causal study of low stakeholder engagement in healthcare simulation projects. *Journal of the Operational Research Society*. 66 (3).

Keehan, S.P., Poisal, J.A., Cuckler, G.A., Sisko, A.M., Smith, S.D., Madison, A.J., Stone, D.A., Wolfe, C.J., Lizonitz, J.M. (2016). National health expenditure projections, 2015–25: Economy, prices, and aging expected to shape spending and enrolment. *Health Affairs*. 35 (8).

Koizumi, N., Kuno, E., Smith, T. E. (2005). Modeling patient flows using a queuing network with blocking. *Health care management science*. 8 (1).

Kolker, A. (2008). Process modelling of emergency department patient flow: Effect of patient length of stay on ED diversion. *Journal of Medical Systems*. 32 (5).

Lin, D., Patrick, J., Labeau, F. (2014). Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Health care management science*. 17 (1).

Liu, S.D., Rudd, A., Davie, C. (2011). Hyper acute stroke unit services. *Clinical Medicine*. 11 (3).

Marazzi, A., Paccaud, F., Ruffieux, C., Beguin, C. (1998). Fitting the distributions of length of stay by parametric models. *Medical Care*. 36 (6).

Mayhew L., Smith, D. (2008). Using queuing theory to analyse the government's 4-H completion time target in accident and emergency departments. *Health Care Management Science*. 11 (1).

Monks, T., Pearn, K., Allen, M. (2015). Simulation of stroke care systems. *Proceedings of the 2015 Winter Simulation Conference*.

Monks, T., Worthington, D., Allen, M., Pitt, M., Stein, K., James, M.A. (2016). A modelling tool for capacity planning in acute and community stroke services. *BMC Health Services Research*. 16 (1).

Moreno, L., Aguilar, R. M., Martın, C. A., Piñeiro, J. D., Estevez, J. I., Sigut, J. F., Sanchez, J.L., Jimenez, V. I. (1999). Patient-centered simulation tool for aiding in hospital management. *Simulation practice and theory*. 7 (4).

Morris, S., Hunter, R.M., Ramsay, A., Boaden, R., McKevitt, C., Perry, C., Pursani, N., Rudd, A.G., Schwamm, L.H., Turner, S.J., Tyrrell, P.J., Wolfe, C., Fulop, N.J. (2014). Impact of centralising acute stroke services in English metropolitan areas on mortality and length of hospital stay: difference-in-difference analysis. *British Medical Journal*. 349 (1).

NHS England (2014). Five year forward view. https://www.england.nhs.uk/five-year-forward-view/

NHS England (2018). Delayed Transfers of Care Data 2017-18. https://www.england.nhs.uk/statistics/statistical-work-areas/delayed-transfers-of-care/delayed-transfers-of-care-data-2017-18/

Osorio, C., Bierlaire, M. (2009). An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research*. 196 (3).

Pitt, M. (1997). A generalised simulation system to support strategic resource planning in healthcare. Proceedings of the 29th winter conference on simulation.

Ramsay, A., Morris, S., Hoffman, A., Hunter, R.M., Boaden, R., McKevitt, C., Perry, C., Pursani, N., Rudd, A.G., Turner, S.J., Tyrrell, P.J., Wolfe, C., Fulop, N.J. (2015). Effects of centralising acute stroke services on stroke care provision in two large metropolitan areas in England. *Stroke*. 46 (8).

Rashwan, W., Abo-Hamad, W., Arisha, A. (2015). A system dynamics view of the acute bed blockage problem in the Irish healthcare system. *European Journal of Operational Research*. 247 (1).

Rodrigues, F., Zaric, G. S., Stanford, D. A. (2017). Discrete event simulation model for planning Level 2 "step-down" bed needs using NEMS. *Operations Research for Health Care*. 17.

The Health Foundation (2016). The challenge and potential of whole system flow. http://www.health.org.uk/publication/challenge-and-potential-whole-system-flow

The Health Foundation (2018). Establishing an NHS-R community to exploit the power of R for the NHS. https://www.health.org.uk/programmes/advancing-applied-analytics/projects/establishing-nhs-r-community-exploit-power-r-nhs

Vasilakis, C., El-Darzi, E. (2001). A simulation study of the winter bed crisis. *Health Care Management Science*. 4 (1).

Vasilakis, C., Marshall, A.H. (2005). Modelling nationwide hospital length of stay: opening the black box. *Journal of the operational Research Society*. 56 (7).

Vasilakis, C., Pagel, C., Gallivan, S., Richards, D., Weaver, A., Utley, M. (2013). Modelling toolkit to assist with introducing a stepped care system design in mental health care. *Journal of the Operational Research Society*. 64 (7).

Vidyarthi, N., Kuzgunkaya, O. (2015). The impact of directed choice on the design of preventive healthcare facility network under congestion. *Health Care Management Science*. 18 (4).

World Health Organisation (2015). Improving health system efficiency. http://www.who.int/health_financing/documents/synthesis_report/en/

Wood, R.M., Griffiths, J.D., Williams, J.E., Brouwers, J. (2014a). Optimising resource management in neurorehabilitation. *NeuroRehabilitation*. 35 (2).

Wood, R.M., Egan, J.R., Hall, I.M. (2014b). A dose and time response Markov model for the in-host dynamics of infection with intracellular bacteria following inhalation: with application to Francisella tularensis. *Journal of The Royal Society Interface*. 11 (95).

Appendix A: Assessment of literature against outlined modelling requirements

A literature review has been performed which has included forward and backward citation searches from the major review papers using Web of Science and PubMed. Publications have been shortlisted on the criteria that they stochastically appreciate patient flow around a network of service points. Shortlisted papers are then reviewed against criteria based on the modelling requirements outlined in Section 1.1. In assessing flexibility the genericity scale introduced in Fletcher and Worthington, 2009 has been applied. Values range from 1 – versatile model across fields and problems – to 4 – specific model for a particular setting and problem.

Table A1 Evaluation of shortlisted papers against required criteria. Column indices measure 1) whether accounts for blocking after service, 2) whether implemented in freely available software, 3) genericity of method, 4) whether delays off the modelled pathway are modelled, 5) whether inputs are intuitive handles (include e.g. arrival rates and LOSs), 6) whether outputs are sufficient for meaningful cost effectiveness appraisal, and 7) whether what-if scenario analysis can be facilitated. Key for columns 1-2 and 4-8 is Y (yes), N (no), P (partially), and U (unclear). Scale for column 3 is based on Fletcher and Worthington, 2009.

	1	2	3	4	5	6	7
Bayer et al, 2010	U	N	4	N	Y	Y	Y
Blasak et al, 2003	N	N	4	N	U	P	Y
Brailsford et al, 2013	U	N	2	U	Y	Y	Y
Bretthauer et al, 2011	Y	Y	4	N	Y	Y	Y
Cahill & Render, 1999	N	N	4	N	Y	Y	Y
Cochran & Bharti, 2006	Y	N	4	N	Y	Y	Y
Cordeaux et al, 2011	U	N	3	N	Y	Y	Y
De Bruin et al, 2007	Y	Y	4	N	Y	Y	Y
El-Darzi et al, 1998	Y	N	4	N	Y	Y	Y
Griffin et al, 2012	Y	N	4	N	Y	Y	Y
Harper, 2002	U	Y	2	N	Y	Y	Y
Harrison, 2001	N	Y	3	N	Y	P	Y
Koizumi et al, 2005	Y	N	4	N	Y	Y	Y
Lin et al, 2014	Y	Y	4	N	Y	P	Y
Monks et al, 2016	N	N	4	N	Y	Y	Y
Moreno et al, 1999	U	N	4	N	Y	Y	Y
Osorio & Bierlaire, 2009	Y	Y	2	N	Y	Y	Y
Pitt, 1997	U	N	2	N	Y	Y	Y
Rodrigues et al, 2017	N	N	4	N	Y	Y	Y
Vasilakis & El-Darzi, 2001	P	N	3	Y	Y	Y	Y
Vasilakis et al, 2013	N	Y	4	N	P	P	Y

Appendix B: Event types for state space transitions

There are three types of event which precipitate movement around the state space, and for which all patient movements into, around, and out of the pathway are associated. These are defined explicitly within the model code referred to in Section 3 and are summarised as follows.

A1. External arrival
The first is an external arrival into a service point *f*, which occurs at rate λ_f. If one of the c_f service channels is available at the service point concerned then the new state occupied is one for which $\eta_{f,1}$ is increased by one. Otherwise the new state occupied is such that ξ_f is increased by one.

A2. Active LOS phase completion
The second type of event is an active LOS service phase completion in a service point *f*, which occurs at rate μ_f. If the phase, *a*, completed is not the final active LOS phase (i.e. $a < k_f$) then the patient remains in their active LOS and the new state occupied is simply where $\eta_{f,a}$ is reduced by one and $\eta_{f,a+1}$ is increased by one.

Otherwise ($a = k_f$), the patient is selected a discharge destination according to the set of probabilities attached to the possible downstream routes out, Φ_f . If the destination, *d*, is not on the modelled pathway (i.e. $d \subset \bar{F}$)

then the patient will either wait if there is an associated modelled delay (i.e. $p_d \geq 1$) or be discharged and transferred (if $p_d = 0$). In the former, the new state occupied will be one in which $\eta_{f,a}$ is reduced by one and $\psi_{f,d,1}$ is increased by one. In the latter, the new state occupied will be one in which $\eta_{f,a}$ is reduced by one (*).

If the discharge destination is on the modelled pathway (i.e. $d \subset F$) and there is no capacity at it (i.e. $\eta + \omega + \psi = c_d$) then the patient remains at the service point and the new state occupied has $\eta_{f,a}$ reduced by one and $\omega_{f,d}$ increased by one. If there is capacity at the destination ($\eta + \omega + \psi < c_d$) then the patient is discharged and transferred and the new state occupied has $\eta_{f,a}$ reduced by one and $\eta_{d,1}$ increased by one (*).

* Events involving patient discharge must also appreciate the potential consequential upstream impacts associated with the vacating of a service channel. If there are no patients waiting for service point *f* at upstream service points then there are no further conditions made of the new state occupied. If, however, there are patients waiting (i.e. if $\sum_{u \in F} \omega_{u,f} > 0$) then their discharge, transfer, and admission to service point *f* need to be appreciated. First the new state occupied has $\eta_{f,1}$ increased by one. Secondly, $\omega_{\bar{u},f}$ is reduced by one where $\bar{u}$ is selected from *F* as the service point which has the highest waiting list for *f* (i.e. $\bar{u} = \arg\max_u \omega_{u,f}$). Finally, such discharge from $\bar{u}$ to *f* would of course make available a service channel in $\bar{u}$, and thus this process would be recursively repeated until there are no such instances.

A3. External delay phase completion
The third type of event is a phase completion within a delay component at service point *f* to an external destination *d*, which occurs at rate γ_d. If the phase, *b*, completed is not the final external delay phase (i.e. $b < p_d$) then the patient remains at service point *f* and the new state occupied is simply where $\psi_{f,d,b}$ is reduced by one and $\psi_{f,d,b+1}$ is increased by one. Otherwise ($b = p_d$), the patient is discharged to the external destination, reducing $\psi_{f,d,b}$ by one and entering the afore-mentioned process * associated with the vacating of a service channel.

Appendix C: Distribution fitting for active LOS

The Akaike Information Criterion (AIC) is a statistical measure which can be used to avoid overfitting in model selection. It works by penalising likelihood by the number of parameters in order to ensure parameters are only included if they significantly improve goodness of fit. It is defined as

$$AIC = 2\kappa - 2\log L$$

where κ is the number of parameters and *L* is the maximum value of the respective likelihood function.

Table C1 details the AIC values for an Erlang distribution fit to active LOS for one to five phases (i.e. $\mu \in \{1,2,\ldots,5\}$), noting that $\kappa = 2$ for $\mu \geq 2$ (since when $\mu = 1$ the two-parameter Erlang distribution reduces to the single-parameter exponential distribution). It can be seen that in all cases the exponential distribution is the optimal choice. That is, there is no statistically significant improvement in goodness of fit associated with the additional parameter of an Erlang distribution. The parameter estimates which optimise the AIC function are given in Table 3.

Note that due to a lack of reliable data for delays to discharge off the modelled pathway, this selection process is not used and instead exponential distributions are assumed around the estimated means (Table 3).

Table C1 AIC for Erlang approximations to active LOS

	Number of Erlang phases				
	1	**2**	**3**	**4**	**5**
Unit X Acute	1,170	1,256	1,401	1,569	1,750
Unit X Rehab	783	819	896	990	1,093
Unit Y Acute	235	244	264	289	317
Unit Y Rehab	346	401	468	540	615
Unit Z	197	228	266	307	349