

‘Ebb and flow by the moon’: can we do better in hospital?

Author: Rupert Negus^A

ABSTRACT

Hospitals are required to manage acutely ill patients. Optimising patient experience with regard to clinical outcomes and patient experience relies on the interaction with healthcare professionals and the manner in which patient cohorts are managed. The key to success is to keep variability to a minimum, both in terms of clinical practice and in terms of how patients are moved through the system; variability in flow increases the likelihood of queues developing, and the development of queues results in delays and has a negative impact on care. This article attempts to illuminate some of the principles which should be considered when designing whole systems. The application of relatively few principles should enable optimal flow of patient cohorts; this in turn will support optimal individual clinical management. As each site is necessarily slightly different, the figures provided are designed to simply illustrate the generic points made in the text.

KEYWORDS: Flow, variability, queue, phasic, emergency admissions

Introduction

After the death of Edmund Ironside in 1016, the Danish King Cnut ruled England for the next nineteen years. His name is popularly linked to the story that in order to display his power over all things, Cnut ordered the tide not to come in: *‘sedile suum in littore maris, cum ascenderet, statui iussit’* (‘he commanded that his chair should be set on the shore, when the tide began to rise’).¹

Cnut knew his feet would get wet, for the story was never intended to demonstrate his arrogance, but rather that he understood that he did not have power over natural phenomena (as succinctly described by Shakespeare’s King Lear (viii 8) in the title quote) and his attempt to demonstrate this to his subordinates. How then should we manage the ‘tide’ of patients moving in and out of our emergency departments (ED) every day? Managing this flow is essential to the smooth running of any acute hospital, a point that seems to be painfully reinforced every winter and will continue to be so until we learn how to

Box 1. Patient-centered principles of the Future Hospital Commission.

- > Patient experience is valued as much as clinical effectiveness: patient experience must be measured, fed back to ward and board level and the findings acted on.
- > Responsibility for each patient’s care is clear and communicated: there must be clear and communicated lines of responsibility for each patient’s care, led by a named consultant working with a (nurse) ward manager. Consultants may fill this role for a period of time on a rotating basis.
- > Patients have effective and timely access to care: time waiting for appointments, tests, hospital admission and moves out of hospital is minimised.
- > Patients do not move wards unless this is necessary for their clinical care: care, including the professionals that deliver it, should come to patients.
- > Robust arrangements for transferring of care are in place: between teams when a patient moves within the hospital and when staff shifts change, and between the hospital and the community.

build systems that can cater appropriately for the phasic nature of the demands for emergency health care.

This article seeks to demonstrate how systems may be designed to accommodate emergency patient flows, even allowing for differences in hospital architecture by applying generic principles through which this may be achieved. These are based upon mathematical models that are now well understood and indeed are already being applied to transformational work in parts of the UK.² In this article I hope to present these principles in a way that is easy to understand and can be related to relevant recommendations made by the Future Hospital Commission.³ Indeed, it is likely that only by understanding how to manage the flow of patient cohorts that the core principles laid out by the Future Hospital Commission can be achieved. Box 1 lists five of these core principles. I hope that by reading this article, the relationship between system design and effective care is more obvious.

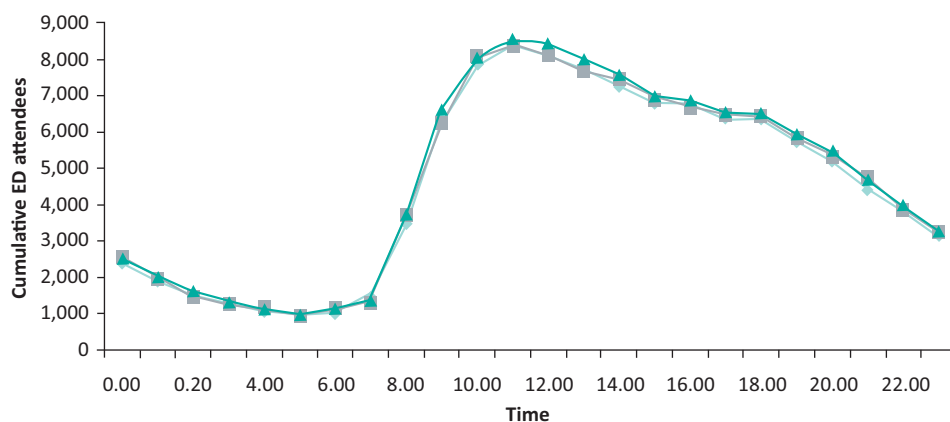
Fundamental principles

Patients ‘flow’ through hospitals, moving through waiting areas to wards or other environments, such as operating

Author: ^Aconsultant physician and gastroenterologist, Royal Free London NHS Foundation Trust, London, UK

Fig 1. Representative graph of cumulative time of arrival for ED attendances at a major London teaching hospital over three consecutive years (represented by the different coloured lines), demonstrating the consistency of arrival times of attendees.

ED = emergency department.



theatres, before supervision of their care is transferred back to the community setting. Multiple moves can occur until the desired outcome is achieved, between each move, patients are held in queues, which are variable in their length and waiting time. Thus a queue for outpatients may be 20 patients long in total, but is expected to be processed over a 3–4 hour period, while the queue for an operating theatre may only consist of 4 patients but may take 12 hours to resolve. Furthermore, the queue for the same process or procedure may vary from day-to-day or hour-to-hour depending on the case mix or the processes used for dealing with the queue. An individual trainee in the ED working on two consecutive days will be faced with two different queues to process, which will take different lengths of time to manage. Conversely, queues of patients waiting for similar diagnostic procedures, such as gastroscopy, may be faced with different operators who in turn may determine how long it takes to resolve the queue.

Principles of queuing theory

The principles behind how queues are generated are fundamental to an understanding of how to optimise patient flow. While any individual queue is likely to be simple, complexity arises due to the number of different queues that form. Erhlang conceived the idea of queuing theory in the early twentieth century to predict performance in telephone exchanges.⁴ The size of a queue is determined by the relationship between the time between arrivals to a queue (A), the size of the job (S) and the number of servers (C , known as Kendall's notation). Thus at traffic lights set to a particular pattern, a queue will not form if the time between cars arriving reflects the time the lights are set to green, but if the pattern of arrivals alters, a queue can form even with the same number of arrivals over a particular period of time.

This analogy can be extended. Consider a toll road in which there are multiple channels through which traffic can flow, each regulated by their own traffic lights. Whether queues form will now be determined not only by the rate of arrival of individual cars but also by the efficiency with which individual drivers interact with the payment system (the size of the job) and the number of available channels (the number

of servers). Inevitably in the real world, the number of servers is limited.

The principles of queuing theory were originally worked out in relation to queues with a Poisson distribution. But in the ED the number of arrivals per hour is not constant, and customarily peaks about midday, displaying a phase-type distribution. Nevertheless the same concepts apply. Whether a queue builds therefore depends on variability in arrival time and the rate at which individuals leave the department. Just as is the case with the toll road, the latter is dependent on the number of servers (ie healthcare professionals) involved in managing the queue as well as familiar variables, such as the nature and severity of illness and frailty, and home circumstances, which will each impact upon the time needed to manage individual patients.

Mismatch between arrivals, the number of servers and the job size results in queues building within the ED, such that by early evening, although the rate of arrival per hour has fallen, the total number of patients in the department reaches a maximum (Fig 1). This in itself generates further inefficiencies as the lack of physical space results in the 'size of the job' for any given condition effectively increasing.

The generation of queues is obviously not unique to the ED; this setting has simply been used as a familiar example. More queues form if the patient is sent for tests, such as an X-ray or scan and arise again when patients require admission and have to wait for specialty opinions. Each queue is called a node and a combination of nodes constitutes a Markov chain. How patients move (flow) overall through a hospital will therefore depend on the interaction of a number of nodes. When a queue forms consistently in time and place within an organisation, it is commonly referred to as a 'bottleneck'; however, the resolution of one bottleneck, usually by introducing more servers may reveal that the next node in the chain is rate limiting.

As can be seen from Fig 1, the rate of arrival of patients in the ED falls to a minimum in the early hours of the morning, thus allowing the queue to resolve overnight and the system to reset for the following day. The familiar pattern of within day admission delay is therefore established and is succinctly described by Allder *et al.*⁵ Other patterns are layered onto this including within week and seasonable variability.

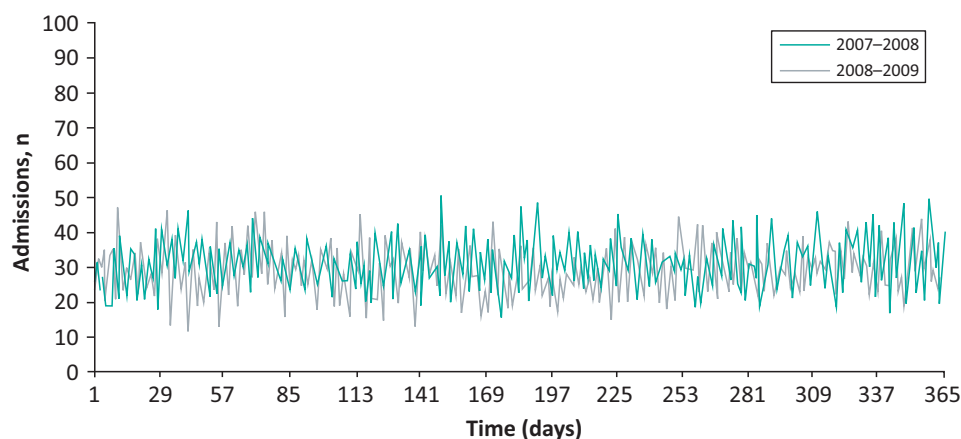


Fig 2. Admissions per day to a London teaching hospital for two consecutive years.

Major flows through hospitals

The major patient flow through a hospital will be familiar to most readers. Numbers vary according to the exact designation of any particular hospital, but for most providing a comprehensive complement of specialties, about 50% of inpatient work is derived from emergency admissions, the majority of which will come in under a medical specialty (including acute and general internal medicine). Approximately 50% will be admitted electively.

Emergency admissions in medicine do show some seasonal variability, and there has been concern about rising numbers, but for hospitals operating where local demographics are stable, the total number of admissions seems to be stable (Fig 2).

By contrast, ED attendances are widely recognised to have risen,^{6,7} as is illustrated by representative data from another hospital in Fig 3 (Dr Andres Martin, consultant in emergency medicine, Royal Free Hospital, London; personal communication). This would imply that the relationship between attendance and admission is fairly flat as is illustrated in Fig 4. Given that the admissions profile is relatively consistent, it is not surprising that the relationship between emergency admissions and performance may also be fairly flat (Fig 5). These numbers and relationships will vary from hospital to hospital, the important message for design rests in

understanding the principles; while the layout of a modern ED may have to take into account rising attendance patterns, it does not necessarily follow that there will need to be an increase in the number of beds required for emergency admissions.

Variability in the interval between arrivals can generate queues. However, while it is traditionally held that greater variability exists in unselected (emergency) patient flows, there is much greater variability in both the arrival of elective admissions and in elective and emergency discharges. Where this has been assessed (Peter Greengross; personal communication), all show greater variability than emergency admissions. Again this variability will contribute to queue formation. Furthermore the interaction between elective and emergency work streams may also have a negative impact on overall performance (Fig 6). The mathematics to describe such an interaction is complex; a useful analogy may be to think of the region where two streams of water meet and the turbulence that is generated at the interface. The relationship between elective activity and overall ED performance is then not surprising, with increasing elective activity being associated with a reduction in ED performance.

Elective admissions obviously impact on bed occupancy and the oft quoted figure of 85% overall bed occupancy being optimal for the efficient running of a hospital reflects this

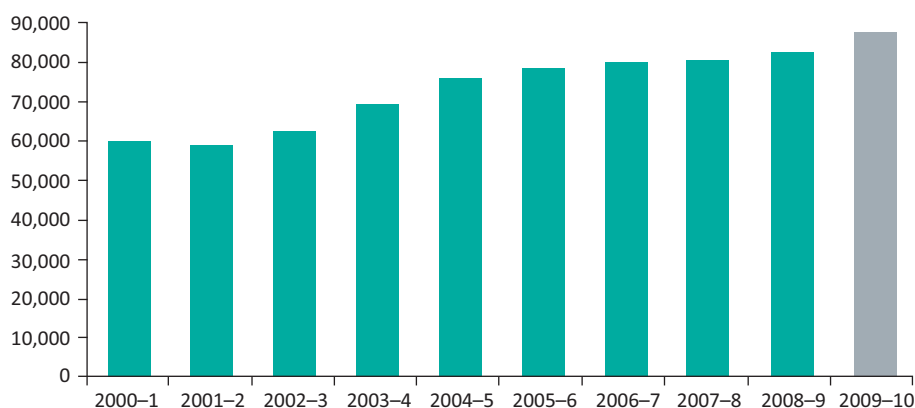
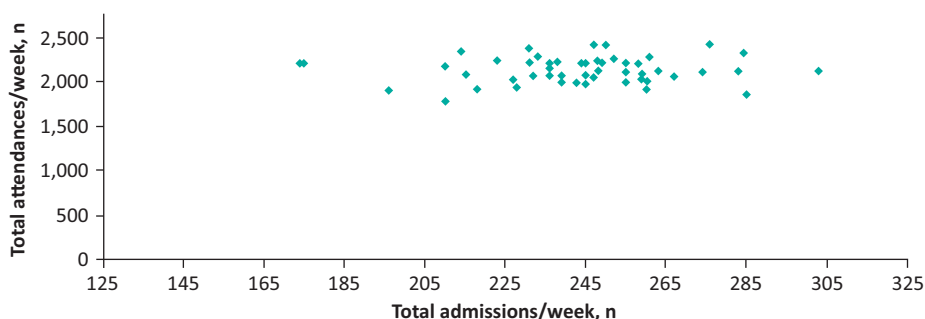


Fig 3. Emergency department attendances between 2000–2010 at another London teaching hospital.

Fig 4. Relationship between ED attendances and admissions on a single site.

ED = emergency department.



thinking,⁸ but it is important to understand that this particular figure only arises as a result of the inherent variability in the system and the flow that is therefore achieved. Furthermore, if this figure is consistent between different hospitals, it lends weight to the argument that the problems they face are essentially the same.

Seasonal variability

I have already alluded to seasonal variability, but comparing Fig 2 with Fig 7 it is apparent that this is less marked in adults than in children.

In adults, while there is undoubtedly an effect, much of the increase in bed occupancy can be attributed to a reduction in the rate of discharge.⁵ For the adult population of inpatients, bed occupancy resets at a new, higher, steady state. The trajectory that links one level of bed occupancy to another is described by an exponential curve that is itself dependent on the shape of the length of stay histogram. Therefore even when the number of admissions falls and the rate of discharge increases, a number of weeks will elapse for this to become apparent in the bed base.

Within-day variability

Within-day patterns of arrivals in the ED are predictable, the maximum rate occurring between 11.30–12.30 and tailing off gently thereafter until midnight, with a characteristic hump at about 4pm, probably due to the arrival of referrals from general practice (Dr Henrietta Hughes; personal communication). As described above, this phenomenon contributes to the generation of queues on a daily basis. There

is also variability in the admissions profile throughout the week, the highest numbers generally occurring on a Monday and Friday (Fig 8).

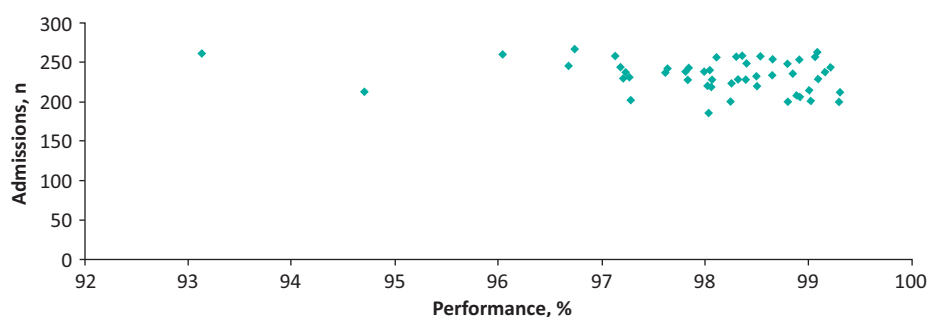
The formation of within-day queues has been used to illustrate the nature of the problem facing patient flow and has been dealt with elsewhere, and explains both within-day and weekly variability in bed occupancy as well as dealing with reasons for seasonal variation.⁵ Within week variability has been discussed above. However, a method of analysing wave forms, known as empiric mode decomposition, offers a tantalising glimpse at other patterns that may influence the overall attendance profile (Fig 9).

It should now be clear that there are many and varied influences on patient movement which contribute to queues building in a variety of settings within the traditional hospital building and at a variety of times daily, weekly and seasonally. Since the admissions profile is reasonably consistent, it should be possible to build systems that will predict the number of emergency admissions to a reasonable tolerance. Indeed the commercially available *Forecaster* system, which takes into account historic admission data and other important factors, such as meteorological data and major sporting events, claims to predict the number of admissions to within approximately 2%, two weeks in advance (Tim Hankey; personal communication). Reliable models can be built and could thereafter be improved in an iterative fashion.

Design principles to optimise flow

When considering how best to use a hospital bed-base, emergency flow is the parameter that is effectively fixed;

Fig 5. Relationship between emergency admissions and performance at the same London teaching hospital.



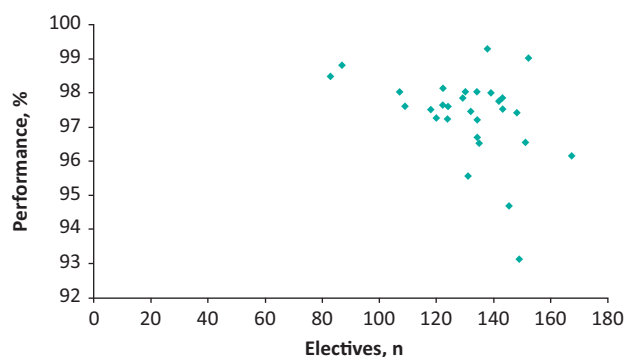


Fig 6. Relationship between performance and the number of elective patients admitted at a London teaching hospital.

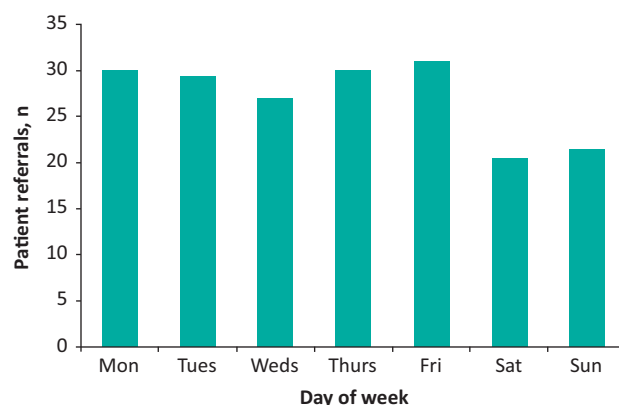


Fig 8. Variation in the median number of ED referrals to the admitting medical team (February–March 2010). ED = emergency department.

optimising the bed-base to accommodate this should be considered first. Initiatives to prevent hospital admission have not yet come to fruition and may be poorly conceived, in that a timely, focused intervention may be more effective than avoiding a visit altogether.

It follows that elective flow must be designed around the requirements for emergency flow, both on a weekly and seasonal basis. There may be room to play with options, such as expanding the bed base at particular times of year, and it might be better if this was aimed at the elective population, since it impacts negatively on emergency flow, rather than the emergency population (the focus usually employed to resolve the issues around bed occupancy that occur particularly in the winter months).

An oft-quoted figure in designing the bed base is that an overall occupancy rate of 85% optimises flow. This figure appears to have merged from Erhlang's studies of telephone operators and is clearly dependent on the rate at which a queue can be resolved, which itself depends upon the variability within the system. In other words, higher bed occupancy could be achieved with less variability and hence higher flows. Another solution might be to engineer the system in such a way as to try and confine 'variability' as much as possible. Diane Pumphrey of the King's Fund talks in terms of 'chains' and 'shops' when thinking about

managing streams of patients (Diane Pumphrey; personal communication). A 'chain' is akin to a conveyor belt, and this concept can be applied to conditions when there is a high degree of predictability about the outcome, for example day surgery for elective hernia repair. This principle can be applied to emergency flows if the degree of diagnostic certainty is high, even in cohorts of patients presenting with high acuity conditions, such as ST elevation myocardial infarction, provided they have a predictable outcome in terms of likely length of hospital stay, complications and mortality. The principle can also be applied to ambulatory emergency care where a set of conditions can be diagnosed and treated rapidly and accurately, and where simple interventions result in a large risk reduction. Conditions where diagnostic uncertainty exists may be better served in the 'shop' setting, where the first priority is to obtain reasonable diagnostic certainty, which in itself may involve exploring various possibilities, and require multiple tests and specialty opinions. Over the past 20 years, developments in diagnostic imaging and laboratory-based tests means achieving such a diagnosis has become more rapid and accurate, allowing more conditions to be dealt with using a 'chain' approach.

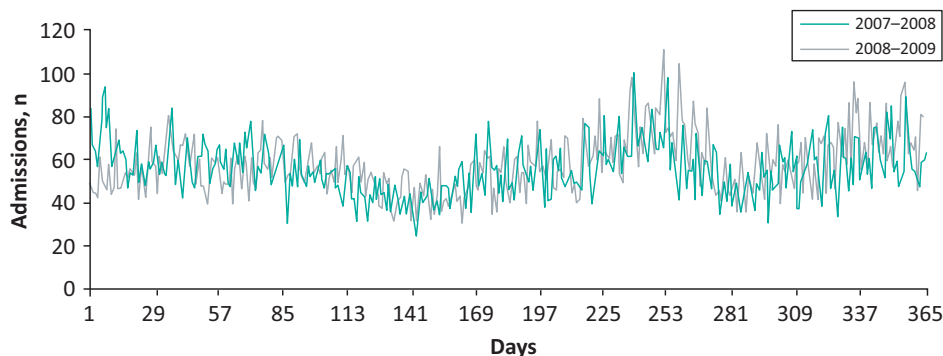


Fig 7. Seasonal variation in paediatric admissions (day 1 is 1 April). There are obvious reductions in admissions corresponding to school holiday periods and increases over the winter months.

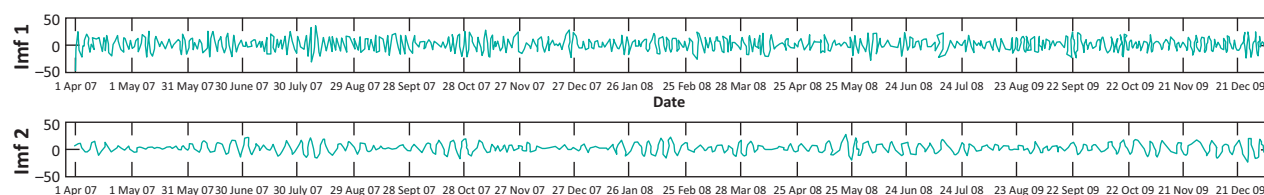


Fig 9. Empiric mode decomposition applied to emergency adult attendances. The first round of decomposition reveals a regular cycle of attendance numbers over a six-week period. The seventh round probably illustrates the longer pattern of seasonal variation running throughout the year.

Common assumptions confirmed and refuted

Before drawing these ideas together to suggest a generic system broadly applicable to any hospital system and which is amenable to predictive modeling, it is worth reiterating some truths and deconstructing some common myths. The medical take is not random, but highly predictable. The elective streams may be largely responsible for impacting negatively on overall hospital performance, as evidenced by the ED performance target. Teams of healthcare workers generally seem to operate at a single 'speed' as inferred from the predictable nature of bed occupancy. If this were not the case it might be expected that teams could 'work' their way out of trouble, but experience does not support this contention. Furthermore if teams slackened off in quiet times, then bed occupancy would be expected to rise, however experience suggests this is not the case; rather, beds become unoccupied for prolonged periods under such circumstances. Because of the variability inherent in patients requiring admission to hospital, it is important to build redundancy into the bed base; this can be focused in a single area, which should make managing variable staffing levels easier. How big this area needs to be depends on the flow through the system, rather than the absolute numbers admitted.

In short, focusing on flow and reducing variability should optimise the system. And indeed one of the key messages in the NHS Wales 1000 Lives Plus national improvement programme is to focus on flow before removing capacity.²

Summary and conclusions

The ideas presented above may be familiar to some readers, while to others appear to be an anathema. However, a hospital functions effectively as a machine and while the presence of brilliant physicians and surgeons is helpful, they are only part of the ingredients needed to run a successful organisation.

A system built around the principles of queuing theory will allow improved patient flow. Returning then to the principles laid out by the Future Hospital, I would contend that this will necessarily enhance patient experience and provide consistently timely care. Patients would be managed in the most appropriate environment within the hospital and unnecessary moves avoided. Minimising moves will minimise the information that can be lost in multiple handovers.

In an era in which evidence-based medicine is rampant, it is important to assert 'absence of evidence is not evidence of absence'. However, to date I am not aware of any systematic attempt to design a hospital with these principals in mind, possibly because while each individual step is simple, complexity arises from the multiplicity of steps involved.

A question that has become more obvious over the past decade, following the introduction of the four-hour target, is what should we be aiming for? Why four hours? Is this simply a reasonable period of time or are there more important effects by providing a specific in terms of the timeliness and quality of patient care. Perhaps the system should be redesigned to optimise patient care in terms of morbidity and mortality rates, readmissions and patient satisfaction, and then the optimal target would simply reveal itself. Working patterns could be geared around outcomes as well, rather than simply assuming that increasing consultant presence improves outcomes. Indeed this does not appear to be the case if the mortality rates from individual hospitals are plotted against the self-reported number of hours of consultant presence derived from the London Safe audit.⁹

Perhaps now is the time to designate an 'experimental hospital' in which, without compromising medical or nursing standards, different flow models could be tested. Indeed, a leaf could be taken out of the Swedish concept of designing a hospital with moveable units, such that different architectural arrangements could also be evaluated. ■

References

- 1 Huntingdon H, Arnold T. *Historia Anglorum. The history of the English from AC 55 to AD 1154: in eight books*. Cambridge: Cambridge University Press, 2012.
- 2 1000 Lives Improvement. *The National Patient Flow Programme. Module two – diagnostics and measurement*. Available online at www.1000livesplus.wales.nhs.uk/sitesplus/documents/1011/flow%20guide%20module%20latest%20version%2027%205%2014%20v3%20FINAL.pdf [Accessed 30 March 2015].
- 3 Future Hospital Commission. *Future hospital: caring for medical patients*. A report from the Future Hospital Commission to the Royal College of Physicians. London: Royal College of Physicians, 2013.
- 4 Sundarapandian V. *Probability, statistics and queueing theory*. Delhi: PHI Learning, 2009.
- 5 Allder S, Silvester K, Walley P. Understanding the current state of patient flow in a hospital. *Clin Med* 2010;10:441–4.

- 6 The King's Fund. *Blog: April 2013*. London: The King's Fund, 2013. Available online at www.kingsfund.org.uk/blog/2013/04 [Accessed 30 March 2015].
- 7 Capewell S. The continuing rise in emergency admissions. *BMJ* 1996;312:991–2.
- 8 Bagust A, Place M, Posnett JW. Dynamics of bed use in accommodating emergency admissions; stochastic simulation model. *BMJ* 1999;319:155–8.
- 9 NHS. *London Health Programmes*. London: NHS. Available online at www.londonhp.nhs.uk [Accessed 30 March 2015].

**Address for correspondence: Dr R Negus, Royal Free London NHS Foundation Trust, Pond Street, London NW3 2QG, UK.
Email: rupert.negus@nhs.net**