Clinical Medicine 2011, Vol 11, No 1: 31–4

Is follow-up capacity the current NHS bottleneck?

Steven Allder, Paul Walley and Kate Silvester

ABSTRACT – Capacity and demand theory suggests that the presence of a queue is not necessarily an indication of a shortage of capacity in a system. It is much more likely that either there is a demand and capacity variation that creates queues or there is a delay designed into the system. A shortage of capacity is only really indicated where a backlog is not stable and continues to grow. In this article, data are taken from one NHS trust that provides evidence for a continually growing backlog for follow-up outpatient services. It is believed that these data are representative of most locations within the NHS in England and therefore suggest an immediate shortage in effective follow-up capacity. To avoid compromise to patient care, the problem will have to be addressed before the situation becomes unmanageable. The paper highlights options to reduce or deflect demand or to increase effective capacity.

KEY WORDS: capacity planning, clinics, demand, queues

Introduction

In two previous articles the challenges concerning the management of bed capacity were highlighted.[1,2] It was asserted that the data did not indicate a real shortage of bed capacity in most NHS trusts because most of the problems associated with bed availability can be explained by the presence of demand and capacity variation. In the case of bed management, better availability can be achieved through a combination of actions to reduce demand and capacity variability and by the identification of the true capacity constraint. In many situations, increasing the number of beds can perversely make the situation worse.

Every system usually has one stage in the process that acts as the capacity constraint, and it is essential for effective capacity management of a system to know where this bottleneck exists.[3] In this article the demand and capacity balance for outpatient follow-up clinics is assessed, to identify if this is the part of the patient journey when flow is most likely to be delayed or where capacity is at its lowest. In particular, there are concerns that priorities and targets set by government have conditioned behaviour such that systems are now designed to minimise the waiting time for patients' treatment to start, but the system is not balanced. A potentially false logic is that any delay to the start of treatment, delays its completion. If the system lacks flow or is unbalanced from a capacity perspective, then pushing people into the system to start their treatment does not necessarily help. There is also the constraint of the capacity that is commissioned by primary care trusts. Experience suggests there is often only a rudimentary understanding of true demand in most NHS systems, and it is common for systems to be working at levels higher than those commissioned.

Steven Allder, consultant neurologist, Department of Neurology, Plymouth Hospitals Trust; **Paul Walley,** associate professor, Warwick Business School; **Kate Silvester,** Inquiry into Flow, Cost and Quality Programme lead, The Health Foundation c/o South Warwickshire General Hospitals NHS Trust, Warwick Hospital

How do we know when we are really short of capacity?

In earlier work, it was suggested that most queues are caused by the presence of demand and capacity variation.[4,5] Figure 1 shows three scenarios where queues behave differently.

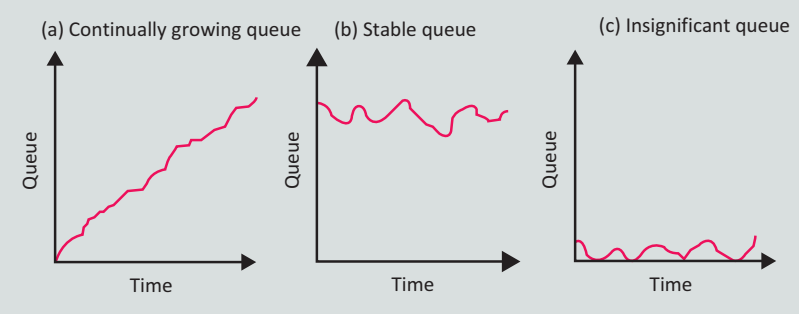
In Fig 1a, the queue appears to grow almost continually, with slight short-term variation in length. This is the one situation where demand clearly exceeds capacity. Each day, more patients are added to the list than are taken off and so the queue increases. By contrast, in Fig 1b, the queue fluctuates around an average and, viewed over a long time, the queue is actually stable. This is where demand and capacity variation create short-term over- and under-supply of capacity, generating queues. This type of queue is known as an 'Erlang' queue. In systems that operate at relatively low levels of utlisation, queues tend not to be problematic. In congested systems that routinely work at high levels of utilisation there is often a tipping point where suddenly queues become more problematic. This is why the 85% bed occupancy rule was originally developed, as a means of managing bed availability. The level of demand and capacity fluctuation affects the tipping point and so queues are more problematic where demand is more uncertain or where capacity is randomly switched on and off more extensively. Figure 1c shows how a queue would behave if demand and capacity variation could be smoothed – the queue reduces greatly without an increase in capacity.

As a general rule, it can therefore be suggested that a shortage of capacity is only indicated when queues continually worsen, and even then existing capacity should not be wasted before more capacity is added to the system.

The theory applied to outpatient clinics

Patient clinics are an ideal example of how capacity and demand theory can be applied to understand and improve a situation. In the majority of situations, there are a number of characteristics that make queues very likely even when there is sufficient capacity.

Fig 1. Queue behaviour.



1 *The system artificially increases demand variation:* first appointment referral requests are often batched between general practitioner and secondary care, creating artificial peaks and troughs in demand.
2 *Capacity variation is introduced into the system:* queues are made worse by the intermittent scheduling of specific clinics on only one or two days per week. Capacity is also often switched off during holiday periods.
3 *Prioritisation increases variation:* one of the known effects of prioritisation systems, such as basic forms of triage, is that it splits demand into subgroups. These subgroups experience relatively higher net demand variation, increasing the average wait in the system.
4 *Subspecialties generate increases in variation:* any form of ring-fencing, such as the creation of a new specialist clinic, increases delay in a very similar way to that of prioritisation. Flexible systems, where work can be spread more evenly across multiple providers, smoothes demand and reduces queues.
5 *Waiting list initiatives increase variation:* attempts to eliminate queues by adding short-term capacity increase the demand variation and generate temporary surges in demand that feed steadily through the whole system. This unbalances the system and moves delays from one place to another. Often the initiative simply switches the delays from one specialty to another, as shared resources are diverted.

Therefore, to maintain the maximum levels of outpatient availability, certain disciplines need to be adhered to. Any form of ring-fencing is ideally avoided, but this can be difficult when patient choice forces some clinics to be dedicated to individual providers. Similarly, continuity of care can place some restrictions on the flexibility of clinic booking. Even when the opportunities to manage demand and capacity have been maximised, natural random variation will give us a trade-off between access and utilisation. If availability and booking flexibility are maintained, levels of clinic utilisation cannot be maintained at 100%.

The above practices only work if there is sufficient capacity in the system. Once the long-run demand of the system exceeds that of capacity, the queue would be expected to grow continually.

Assessing the current state of follow-up clinics

To illustrate the issues identified, demand, capacity and activity data have been extracted from one department of a case study site, which is a large NHS secondary care trust. These data have been carefully validated. Figure 2 shows a comparison of the commissioned capacity, actual available capacity and used capacity in the case study site for new outpatient attendances. The first observation is that the trust appears to actually fulfil more attendances than it is commissioned to provide. The trust's willingness to provide this capacity is possibly due to the fact that it risks breaching waiting time targets if it does not accept new patients within a reasonable timeframe. There is also a desire to assess patients quickly to minimise the clinical risks associated with long delays to first appointment. As a consequence of this behaviour, the waiting list remains flat.

Figure 3 shows the same data for follow-up clinics. In this case, additions to the lists and commissioned capacity seem to be more balanced, with some occasions where demand exceeds capacity and vice versa. However, the actual attendances achieved by the system seem to be considerably lower than demand and the system looks as if it is not operating at 100% utilisation. The net result is that the total queue length is steadily

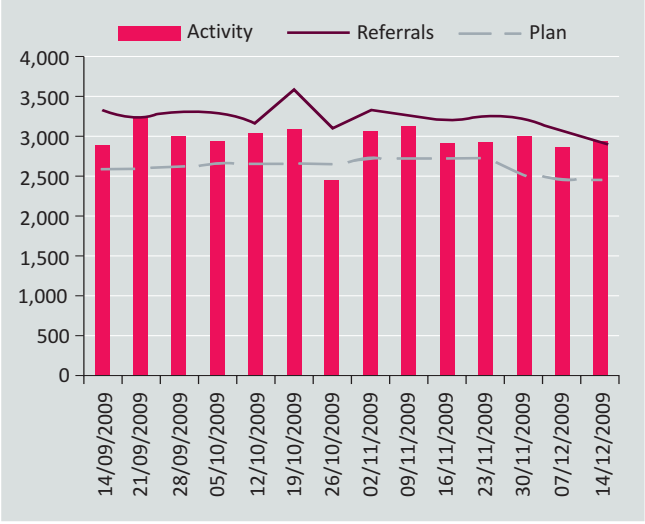


Fig 2. A comparison of first appointment demand, capacity and activity.

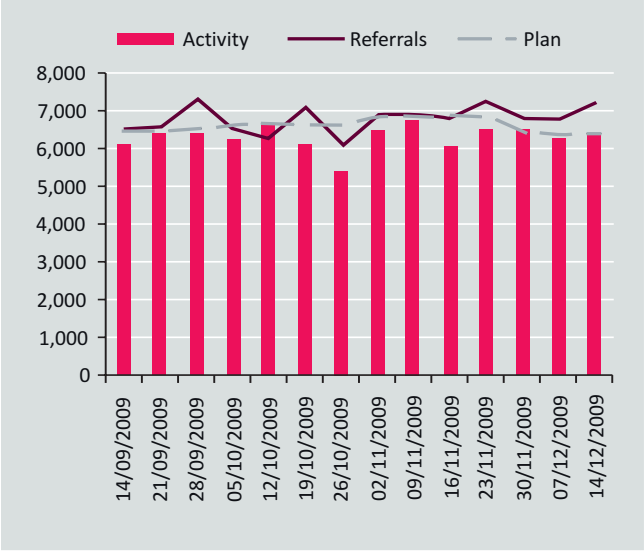


Fig 3. A comparison of follow-up appointment demand, capacity and activity.

growing, resulting in ever-increasing delays for patients waiting for follow-up appointments.

The situation seems to be perverse in that the part of the system with insufficient commissioned capacity seems to have less of a problem in keeping queue length down, whereas the part of the system with less than 100% utilisation fails to fulfil requirements. There are three mechanisms at play here. First, the system naturally prioritises first appointments, irrespective of the commissioned capacity, so that target waiting times are not breached. It partly achieves this by using capacity that could be used for follow-up appointments instead. Secondly, the follow-up capacity has greater demand and capacity variation. The view from the consultant's perspective is that the initial demand received by the trust as first appointments is batched as it goes through the system, resulting in large swings in demand from one week to the next. Additionally, the capacity varies from week to week due to holidays and other closures, creating further short-term mismatch between capacity and demand. Thirdly,

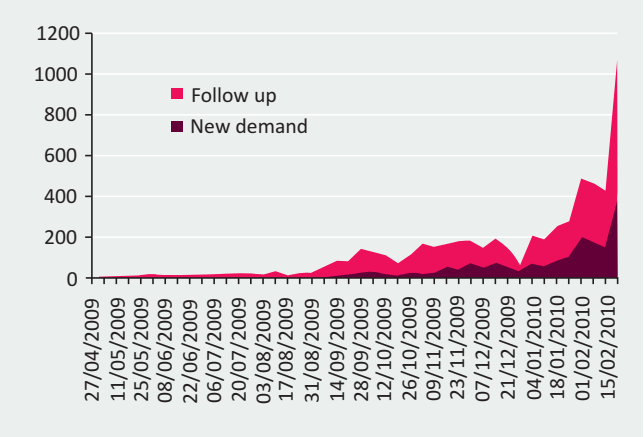


Fig 4. Total outpatient work backlog.

the patients require some degree of choice about the timing of their follow-up appointments. These factors mean that the follow-up clinic system cannot reasonably expect to operate at 100% utilisation and the proportion of unused appointments is influenced as much by the variation and patient choice as it would be by 'better' management of clinics.

The lack of understanding of demand and its behaviour is probably a contributing factor to the problems faced by trusts. Contracts with primary care are often based on historical activity statistics, not demand, coupled with legacy budgets. Consequently, capacity is set at what is affordable rather than what is necessary. Follow-up capacity is based on historic ratios of new:old appointment activity which neither reflects true demand or the future possible ratio of new:old demand. For example, it is likely that new:old ratios change due to the better management of patients with long-term conditions. At present, the medical model is to manage the more complex long-term conditions within secondary care. Most capacity planning models no longer allow for multiple follow-up appointments for these groups of patients, resulting in a systematic under-reporting of demand. The system is also distorted by waiting list initiatives which mask the underlying demand trends and create large swings in demand for resources, ensuring that the queues become worse.

The patient experience

Some patients, especially those classed as high priority, will experience a good service. Most patients, however, will be put into an ever-growing queue where the waiting time for follow-up will increase steadily over time. Those patients graded as low priority for follow-up presently may never actually receive their follow-up appointments as higher priority patients will continue to be scheduled sooner. The risk is that de facto the problem is managed by quietly dropping patients from follow-up lists with arguments based around lack of clinical need.

Suggested actions

There are four main changes that need to be made to the system.

1 *Demand-driven capacity planning:* there are several ways in which capacity planning of clinics can be improved. First, commissioners must look at demand and see what capacity is needed to meet that demand, rather than the classic public sector approach of a budget-based resource allocation. Second, the plan must allow for demand and capacity variation, which means planning on the basis of less than 100% utilisation for follow-up clinics. Third, clinic demand is relatively deterministic. In most cases it is possible to understand the relationship between the number of episodes of treatment and the number of clinic slots required to satisfy demand (as well as the approximate timing of the clinics needed). This calculation is not being made, especially for demand from patients with long-term conditions who are

likely to need repeated follow-up appointments, potentially for many years after first entering the system.

2 *No capacity 'lag':* the capacity strategies of trusts must anticipate the areas where demand is likely to grow over the next few years and avoid a 'capacity-lag' strategy, ie they must not wait for a queue to get out of control before reluctantly adding more capacity.

3 *Demand deflection:* it is appropriate to ask whether or not some of the follow-up work can be done in other settings, such as primary care or within the community care system. There will always be a need to retain some capacity for long-term conditions within secondary care, but it does not necessarily have to supply all follow-up capacity.

4 *Managing the existing backlog:* there is already a problem with significant numbers of patients waiting on growing lists without the immediate prospect of being seen in a timely manner. The problem must be addressed, but without resorting to the option of a waiting list initiative. Instead, the necessary appropriate adjustments to existing capacity should be made and the queue allowed to fall at a natural level. Some patients could also be transferred to other sources of supply. In the short term, system utilisation may be higher than normal and this would fall to appropriate levels as the queue is reduced.

Conclusions

Most waits and delays can be greatly reduced by better control of capacity and demand variation. Previous articles have shown how such principles can be applied to bed management. In this article, the same discipline has been applied to outpatient clinic capacity. Despite the rigours of the capacity and demand balance assessment, which would normally find ways to eliminate delays without resorting to more capacity increases, there is strong evidence to suggest a capacity shortage for follow-up outpatient clinics. There are several reasons for the current situation. Primary care trust commissioners and hospital managers have focused attention on the ability of the system to accept new patients with minimum delay. However, this has masked an underlying problem that resultant changes have left the flow of work in the system unbalanced and partly created the problem with follow-up capacity. Work is being pushed into the system, but there is insufficient capacity for it to leave with adequate follow-up. Existing methods of requirements planning for clinic capacity underestimates the impact of patients with long-term care needs. There is usually an allowance of just one clinic slot per new patient when, on average, there is a genuine need for more than. Although the evidence that presented here comes from just one healthcare community, this is likely to be a national-level problem, with such delays occurring widely across the country. It is hoped that this evidence can be used to put into place different methods of capacity planning and act as a catalyst for the development for alternative methods of long-term follow-up care.

References

1 Allder S, Silvester K, Walley P. Managing capacity and demand across the patient's journey. *Clin Med* 2010;10:11–3.
2 Allder S, Silvester K, Walley P. Understanding the current state of patient flow in a hospital. *Clin Med* 2010;10:441–5.
3 Goldratt E, Cox J. *The goal. A process of ongoing improvement.* Croton-on-Hudson, NJ: North River Press, 1984.
4 Silvester K, Lendon R, Bevan H, Steyn R, Walley P. Reducing waiting times in the NHS: is lack of capacity the problem? *Clinician in Management* 2004;12:105–11.
5 Walley P, Silvester K Steyn R. Managing variation in demand: lessons from the UK National Health Service. *J Health Manag* 2006;51:309–20.

Address for correspondence: Dr S Allder, Department of Neurology, Plymouth Hospitals NHS Trust, Derriford Hospital, Derriford Road, Plymouth PL6 8DH.
Email: steven.allder@phnt.swest.nhs.uk