

# Why Do We Have a Waiting List?

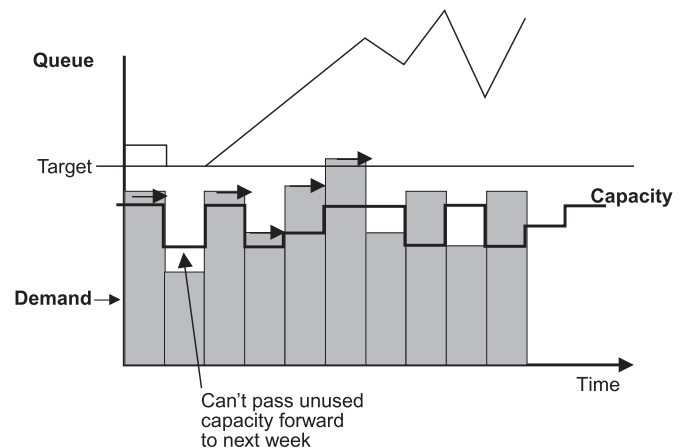
KATE SILVESTER

*Keywords: Waiting lists; Management; Modernisation*

There are several reasons for having a waiting list. First of all we believe that the demand for healthcare exceeds our capacity to provide it. This feeds an irrational fear that, if there was no waiting list, every one would want an operation everyday just to keep us doctors busy. So the queue is 'used' to control the demand and moderate the threshold for providing care. Furthermore, intuitively we know that the demand for healthcare isn't constant – exactly the same number of patients with precisely the same case mix do not fall ill everyday. So we keep a queue to keep our theatre sessions busy. In other words, we keep patients waiting at our convenience, not theirs. However let us challenge the underlying paradigm. If there wasn't a queue, would everyone want every operation possible everyday? Is the demand truly out-stripping our ability to supply it?

As WE Deming so famously said, "every system is perfectly designed to achieve the results it gets" and we have, unwittingly, perfectly designed the systems to have a queue. Firstly, the planning system is not based on the demand (i.e., the additions to the waiting list), but it is based on the average past activity (e.g., last year's average activity +/- 'a bit'). Second, if we plan the average capacity over the year to match the average demand for the year, we will, unwittingly, cause the waiting list. How come?

If we look at the demand for a service, in our case the additions to the waiting list for day case procedures, the demand varies every day. In addition, the case mix changes too. To complicate the system further, the case mix by speciality and then by specific consultant will vary too. When it comes to the capacity, this varies all the time. Each of the specialities dealing with the case mix has a different numbers of sessions each week depending on who is available. Bank holiday Mondays cause obvious variations in capacity. So every time the demand (requests that day) are greater than the capacity that day, then the 'excess' demand will be carried forward as a waiting list or queue (Figure 1). However, if today's demand is less than the available capacity today, we can't pass the unused capacity forward to tomorrow. It is lost. So over time the waiting list persists and serves to give an illusion of high utilisation of future capacity. Since it is impossible to catch up, we have to provide expensive short term capacity, either as 'good will' and overtime in overrunning sessions, waiting list initiatives or private sector provision to scrape the waiting list back to the target waiting time. In addition, to try and protect the vulnerable patients from the queue, we create sub-queues of 'urgent or 'routine' patients.



**Figure 1** The consequences of the mismatch between variations in demand and capacity. Even if the average demand is equal to the average capacity, the consequences of the mismatch between variations in demand and capacity will be a queue. Based on Silvester et al (2004)<sup>3</sup>, with permission

Overall the system becomes impossible to manage, due to the mismatch between the varying demand and types of demand and the capacity and all the sub-divisions of capacity. In addition to the huge overhead and 'non value adding' costs of the managers and clinical time managing the queue, we create risk and frustration for patients and staff alike. Wouldn't we be better off simplifying the system by providing a bit more capacity or slack to cope with the variation in demand? Is there a more rational way of simplifying the capacity? Overall would the return (income) for the overall cost be better if we didn't carry all the overhead?

So how much more capacity do we need? This depends on the response time required. If there are some patients who will need to be seen quickly and cannot tolerate any wait (e.g., myocardial infarct patients), then to avoid the costs of an extra step of triage, we need to see all the ischaemic patients quickly and design the capacity to meet the peaks in demand and accept the 'waste' that will result. However, if we are running a service where no patient needs to be seen straight away, then it is important to recognise that the relationship between capacity utilisation and the waiting time is not linear. Erlang<sup>4</sup> demonstrated that in a system where there is a varying demand and a fixed capacity (i.e., a telephone exchange), the response time is very good until the

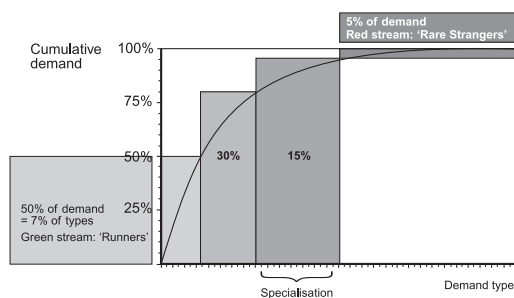
#### Author's Address

KATE SILVESTER Former Coach to the clinical systems engineering programme  
Lean Academy, Heart of England Foundation Trust, Devon House, Birmingham Heartlands  
Hospital, Bordesley Green East, Birmingham, B9 5SS

capacity is 80–85% utilised\*. After this point, the response time increases dramatically. Services that plan to have 95% utilisation of capacity are therefore planning to have a queue beyond any desired waiting time.

So in order to design a cost efficient, responsive service that is immune to the problem of over-runs, waiting lists initiatives and loss of good will, we need to do the following:

1. Measure the demand and variations in demand on a weekly or daily basis
2. Set the average capacity at a level that is appropriate to the response time our most vulnerable patients need, bearing in mind that if we aim for more than 85% utilisation of the capacity, we will guarantee a waiting time beyond our desired response time
3. Add in the short term capacity required to drain out the waiting list in the desired time
4. We also need to think about reducing the variations in capacity



**Figure 2** A Pareto analysis showing the cumulative frequency distribution of the demand by the types of procedures that patients need. Half of the demand is accounted for by only 7% of the procedures. See text for full details

A Pareto analysis (i.e., a cumulative frequency distribution of the demand [y axis] by the types of procedures that patients need [x axis]) is the way of doing this. In a classic Pareto distribution (Figure 2), 80% of the patients will need 20% of the procedures. Further subdivision by Ian Glenday<sup>2</sup> shows that 50% need just 7% of the procedures. This 7% of procedures should be standard work, with all the 'operators' capable of performing them. This is called the Green stream, in which the variation in both the volume and case mix is the least. The required slots should be level scheduled with the same number of slots available every day of every week. The tail end of the Pareto distribution is made up of just 5% of patients who require a huge range of rarer and less predictable procedures. This is the Red stream and should be planned for outside the mainstream and on demand. In the

#### \* Editor's note

Agner Krarup Erlang (1878–1929) was a Danish mathematician and telephone engineer. He applied the theory of probability to his local village telephone exchange to work out the fraction of users trying to place calls outside of the village who would have to wait because all of the 'phone lines were in use. He subsequently developed formulae to determine the optimal number of telephone circuits and telephone operators required to reduce congestion and waiting times to acceptable limits. The Erlang is now internationally used as the unit of telephone traffic flow and his theories are still used by all modern telephone systems. The same principles are commonly used to determine the design and staffing of call centres, which need to provide a service capable of meeting typical demand without allowing excessive queues to form, and are equally applicable to any service with a randomly varying demand, such as the NHS.

middle of the Pareto distribution is another 30% of patients (Yellow stream) who need a larger range of procedures which all the consultant staff should be capable of performing. Only the 15% of patients requiring the wider range of procedures beyond that (Blue stream) will require a specialist, as the volumes here are too small for everyone to be expert.

Hence by rationalising the capacity by the Pareto of the demand, we can start to level the capacity and schedule the work so that we can meet the demand. First of all, we need to think about abolishing the artificial batch quantity called a session as it bears no relationship to the demand. Once the schedule has been defined by the number of slots required for the Green, Yellow, Blue and Red streams of work in each speciality, it is each speciality's responsibility to supply the skills needed to meet this schedule. However, this will mean team working, standardisation, cross training and cooperation between, what are now, autonomous individuals.

If we are in any doubt that we have more than enough capacity now to meet the demand reliably, it is worth spending a day just watching an operating theatre at work. This reveals the amount of waste that is currently caused by an un-level schedule that is being constantly adjusted to meet the mismatches between the demand and capacity, the poor processes, uncertainty and the resulting confusion. Abolishing waiting lists is not about hitting an 'irrational government target' and improving efficiency per se, but about reducing stress and improving safety.

## References

1. Erlang AK. The theory and probabilities of telephone conversations. *Nyt Tidsskrift for Matematik B* 1909;**20**
2. Glenday IF. *Breaking through to flow: banish firefighting and produce to customer demand*. Lean Enterprise Academy, 2005.
3. 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(3)**:105–11.

#### Other useful sources:

- <http://www.leanuk.org>
- [www.steyn.org.uk](http://www.steyn.org.uk)