NHS Improvement

ACT Academy

Online library of Quality, Service Improvement and Redesign tools

Demand and capacity – a comprehensive guide

collaboration trust respect innovation courage compassion

Demand and capacity – a comprehensive guide

What is it?

In order to maximise patient flow through a healthcare system, you need to look at the entire patient process. This guide helps you to understand the demand and capacity of a system and what you can do if there is a mismatch between demand and capacity that has resulted in a backlog (waiting list).

Analysing the demand and capacity within a service will enable improvements to be made that smooth the flow of patients through the system and help to create a better patient and staff experience of the healthcare process.

When to use it

Delays within systems occur when flow into the system is greater than flow out of the system. When this happens, you need to undertake a demand and capacity analysis. If you want to know how well services are performing, you need to have reliable measures for demand, capacity, activity and backlog in place. Having such measures is good management practice and should be routinely and systematically carried out.

By analysing the patterns that emerge from the data collected, you can start to better match demand and capacity, which will ultimately help to reduce and remove the backlog or waiting list.

How to use it

1. Start by defining demand, capacity, activity and backlog for the service you are focusing on.

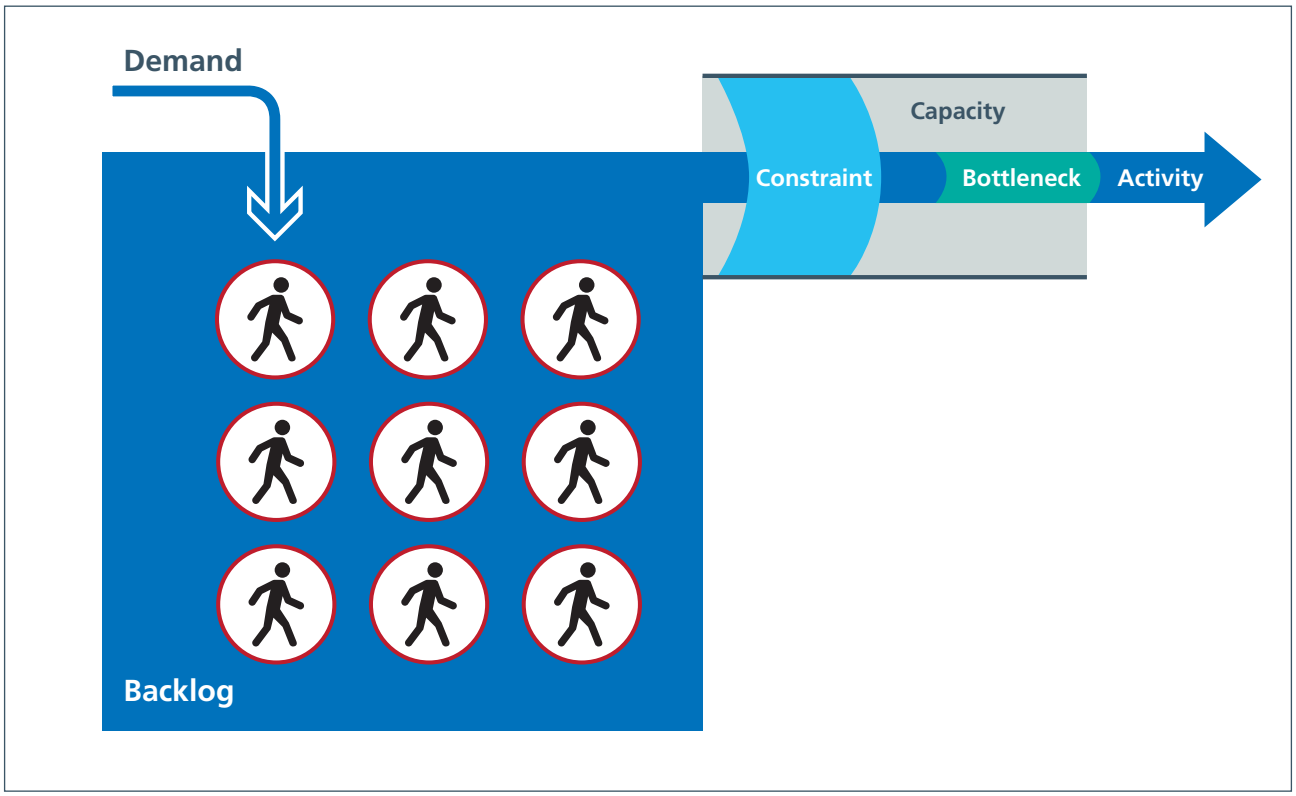
1.1 **Demand** – all the requests/referrals coming in from all sources and what resources they need (equipment time, staff time, room time) to be dealt with. It is easy to miss some demand, especially if you are looking at a system that has a finite capacity, eg for an intensive care unit, the capacity is limited to the number of beds available and so when demand exceeds the number of beds available, something else must happen to the patient such as them being transferred to another hospital, or cared for on a ward with support from a critical care outreach team. It is very important that any demand that is not possible to meet is captured to provide an accurate reflection of true demand.

1.2 **Capacity** – the resources available to do work – for example, the number of pieces of equipment available multiplied by the hours of staff time available to run it. Again it is important to look at actual capacity rather than 'theoretical' capacity. Theoretical capacity can be calculated by looking at staff rotas, etc but there is always a mismatch between theoretical capacity and actual capacity due to lots of factors that may have not been taken into account, eg staff annual leave, study leave, sickness, time out for urgent meetings, etc.

1.3 **Activity** – refers to all the work done. This does not necessarily reflect capacity or demand on a day-to-day basis. The activity or work done on a Monday may be the result of some of Monday's demand (ie emergency) and some carried over from the previous weeks. Activity data is usually more easily available than capacity or demand data, but it should not be used instead of true capacity or demand data.

1.4 **Backlog** – previous demand that has not yet been dealt with, showing itself as a backlog of work or a waiting list. For example, in a community setting, backlog includes all patients waiting to be assessed by the occupational therapist – both those on waiting lists and new patients in the system.

Figure 1: Demand and capacity



2. Map out the processes or patient pathways at a high level (see **process mapping**).

3. Identify the steps or parts of the process where there are the longest delays for patients. A bottleneck is any part of the system where patient flow is obstructed causing waits and delays. It interrupts the natural flow and hinders movement along the care pathway. However, there is usually something that is the actual cause of the bottleneck and this is the constraint.

4. Map this part of the process in more detail so that you really understand what is going on. Map to the level of what one person does, in one place, with one piece of equipment, at one time. The **process templates** tool will help you at this stage.

5. Look carefully for the true constraint. The constraint is often a lack of availability of a specific skill or piece of equipment (eg decontamination machine, CT scanner or specific surgical skills). Waiting lists or backlogs occur before the constraint in the patient journey and clear after the patient has gone past the stage with the constraint. **Process templates** identify the constraint visually.

6. Keep asking 'why?' to try and discover the real reason for the delay (see **root cause analysis using five whys**). For example, the clinic always overruns and patients have to wait for a long time. Why? Because the consultant does not have time to see all her patients in clinic. Why? Because she has to see everyone who attends (including first visit assessments and follow up patients). Why? It is what she has always done.

7. Building on your detailed map, move on to measurement of demand, capacity, backlog and activity. For comparison purposes, they should all be measured in the same units for the same period of time ie hourly, over a 24 hour period, weekly or monthly. It is also important to compare the four measures on a single graph.

How to measure demand

Predicting demand is difficult because there is always variation within the demand data ie there will not be a predictable number of patients referred to a service that is the same every day. Because demand data is difficult to predict, historical activity data is frequently used in its place. However, activity data only shows the number of patients seen or the number of procedures carried out on a specific day so can only be a reflection of the supply of services at that time rather than the true demand.

To measure demand at the bottleneck step, multiply the number of patients referred by the time in minutes it takes to process a patient. Include all demand sources, eg orthopaedics, physiotherapy referrals, pain clinic and direct orthopaedic referrals in order to produce a complete picture of the true demand.

For example, four referrals multiplied by a consultation time of 45 minutes each amounts to 180 minutes (three hours) of demand each day. The illustration below shows how to measure demand.

It is important to plot raw demand data on a graph so that you are aware of the variation within the demand. For example, if a consultation time is an average of 45 minutes and there is a lot of variation for individual patients, it will affect how you need to plan a service. Great care needs to be taken whenever an 'average' is used within a demand and capacity analysis.

Figure 2: How to measure demand

Type of software	Minutes taken to complete task	Requests Mon	Mon total	Requests Tues	Tues total	Requests Wed	Wed total	Requests Thurs	Thurs total	Requests Fri	Fri total	Total requests	Total demand (minutes)	Total demand (hours)
Endoscopy	30	2	60	4	120	5	150	6	180	1	30	18	540	9.0
Colonoscopy	45	4	180	5	225	6	270	7	315	8	360	30	1350	22.6
New consultation	30	2	60	7	210	5	150	3	90	2	60	19	570	9.5
Follow up consultation	20	1	20	3	60	5	100	6	120	4	80	19	380	6.3
CT head	20	7	140	2	40	4	80	1	20	5	100	19	380	6.3
MRI knee	20	4	80	3	60	4	80	4	80	9	180	24	480	8.0
Total	0	20	540	24	715	29	830	27	805	29	810	129	3700	61.7

How to measure capacity

Multiply the number of pieces of equipment by the time in minutes available from people with the necessary skills to use them. For example, two treatment machines multiplied by 480 minutes (eight hours) of session time amounts to 960 minutes (16 hours) of capacity each day. Make sure your calculation takes into account time off for staff breaks, equipment maintenance, etc.

You can then convert capacity into the number of patients that could be seen. So, if a patient takes 20 minutes to process, then the capacity is 960/20, which equates to 48 patients.

Figure 3: How to measure capacity

Name of operator	Mins available (Mon)	Mins available (Tues)	Mins available (Wed)	Mins available (Thurs)	Mins available (Fri)	Mins available (Sat)	Mins available (Sun)	Total
A	240	0	180	0	0	60	0	480
B	0	240	240	0	60	120	0	660
C	180	0	300	0	0	0	240	720
D	0	300	0	300	320	180	0	1100
E	240	0	0	420	0	0	300	960
Daily capacity	**660**	**540**	**720**	**720**	**380**	**360**	**540**	**3920**

Ensure that you measure all available capacity. Many people do lots of different things, so make sure you measure any hidden capacity. For example, if you want to calculate the capacity for a pharmacist dispensing or preparing chemical substances, you would need to know all the activities they currently do and understand the proportion of their time devoted to each task.

Determining the true capacity of a system is often easier than predicting true demand. There are four steps:

i. Determine the overall supply of the service – how much capacity is available, for example minutes in outpatient clinic time?

ii. Consider how supply changes over differing weeks and months (eg staff leave) – it is important to understand actual capacity as opposed to potential capacity and then to look at ways of bringing the two closer together (eg co-ordinating consultant leave may result in fewer clinic cancellations)

iii. Identify how the supply is provided. Can it be provided in shorter time periods? Often services work in 'batches' with, for example, patients waiting for a specific clinic. It can be much more efficient if supply is deployed evenly against demand because the closer demand and supply can be matched, the betterthe system will run

iv. Is the service providing what is really required to meet the patient's needs? For example, the provision of radiology services for the management of DVT may be provided in a more efficient way.

How to measure activity

Multiply the number of patients processed through the bottleneck by the time in minutes it took to process each patient.

For example, 100 patients processed multiplied by 20 minutes each equals 2,000 minutes (33.3 hours) of work done each day.

Warning: measures of activity can be misleading as they do not necessarily reflect demand or capacity. For example, activity in June may well include demand carried over from May, April or even March.

In addition, staff may have not been fully utilised. They may have been kept waiting for the patient, specialised pieces of equipment or test results.

Figure 4: How to measure activity

Sum of minutes to complete task	Day					
Type of request	Monday	Tuesday	Wednesday	Thursday	Friday	Grand total
Colonoscopy	160	180	250	160	90	840
CT	150	70	50	125	40	435
Endoscopy	140	120	200	120	90	670
MRI	90	75	70	150	60	445
Grand total	**540**	**445**	**570**	**555**	**280**	**2390**

How to measure the backlog

Multiply the number of patients waiting by the time in minutes it will take to process a patient through the bottleneck.

For example, 100 patients on the waiting list multiplied by 20 minute treatment time each equals a 2,000-minute (33.3 hours) backlog.

Ensure that you don't count the same patient more than once. There may be patients on waiting lists at different parts of the same process, eg patients requiring radiotherapy treatment can be on waiting lists (or backlogs of work) for their pre-treatment, planning and simulation at the same time. Only count them in the earliest stage to avoid recounting them later in the process. In the radiotherapy example given, it will be at the planning stage.

Figure 5: How to calculate backlog

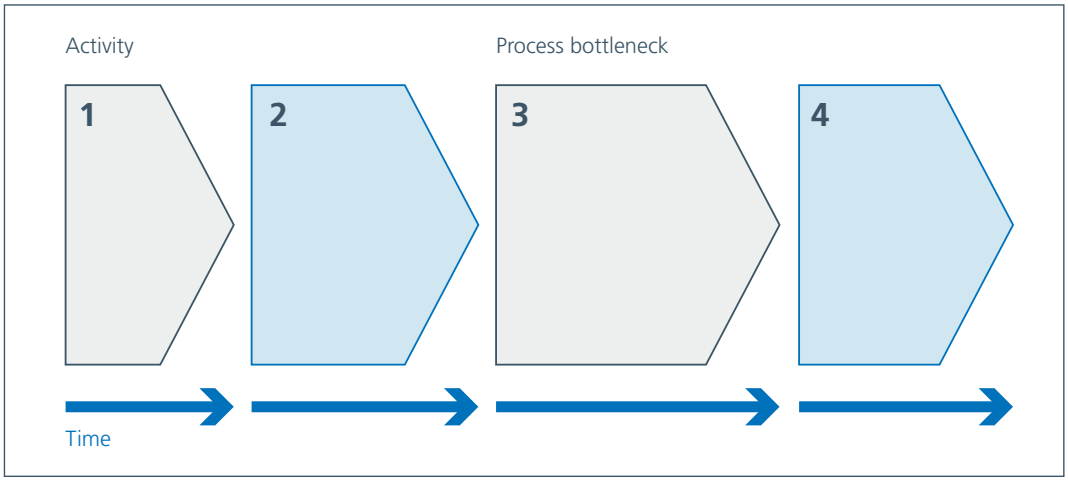
Type of request	Number waiting	Minutes taken to complete task	Minutes taken to clear backlog
CT head	3	20	60
MRI knee	15	20	300
Endoscopy	4	30	120
Total backlog	22	70	480

Identifying a bottleneck

A bottleneck is any part of the system where patient flow is obstructed causing waits and delays. Bottlenecks determine the pace at which the whole service can operate. They have the smallest capacity relative to demand. There are two different types of bottleneck.

- **Process bottlenecks** – the stage in a process that takes the longest time to complete, often referred to as the rate limiting step or task in a process.

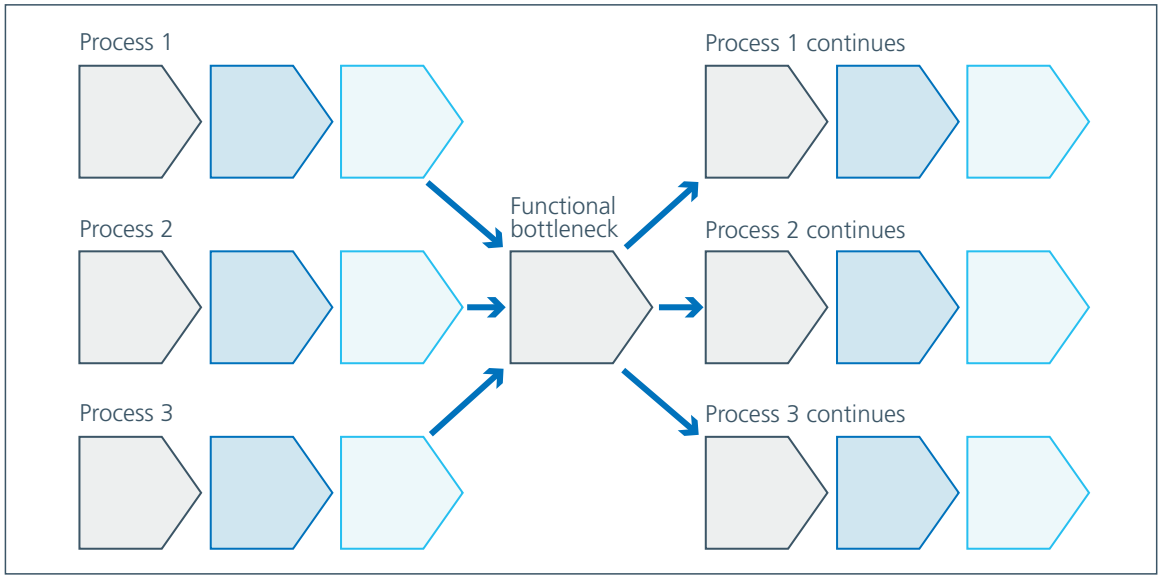
Figure 6: Process bottleneck



- **Functional bottlenecks** – caused by services that have to cope with demand from several sources. Radiology, pathology, radiotherapy and physiotherapy are often functional bottlenecks. They cause waits and delays for patients because one process shares a function with other processes. Staff can be functional bottlenecks as they have a number of different demands on their time, eg a surgeon may be called to theatre when they are also needed to run an outpatient clinic.

This type of bottleneck causes a disruption to the flow of all patient processes. They act like a set of traffic lights, stopping the flow of patients in one process while allowing patients in another process to flow unhindered.

Figure 7: Functional bottleneck



Attempts at improving services will not deliver the necessary improvements if the bottlenecks are not tackled. Any service improvement is unlikely to succeed because the patient will be accelerated through part of the process, only to be halted further along the pathway by the bottleneck.

Once the flow into a service is known on a day-by-day and week-by-week basis, you need to manage capacity to match it, so that flow out of the service is the same. Once you reach this equilibrium, you can work to reduce the backlog and eliminate it.

Your next aim is to ensure that demand and supply remain in equilibrium. This requires you to match and manage supply and demand on a daily basis. This way, the clinical team can avoid backlogs of work building up and the restriction of patient access. It is important to flex capacity as much as possible to meet demand.

Examples

A GP practice situated in an area with some challenging health problems had an average waiting time of 4.79 days to see a GP. The practice reviewed its demand (the number of appointments requested daily) and its capacity (the number of appointments available daily).

This information allowed the practice to change the appointment system to match demand and to introduce different ways of accessing care, eg telephone consultations and access to repeat prescriptions.

Based on the demand and capacity information, a skill mix approach was introduced to ensure patients were seeing the most appropriate member of the healthcare team. The practice was able to reduce the waiting time to 0.32 days – an improvement of 93%.

What next?

Think about demand:

- Continually measure, plot and display demand data.
- Should we see all these patients? Think about implementing protocols to ensure more appropriate referrals.
- Who is the most appropriate professional to see the people referred? Consider alternative ways of working.
- Can the patient pathway or the process at the bottleneck be streamlined? (Do we need to do all these steps?)
- Reduce waiting lists – reduce the demands they create.

Think about capacity:

- Continually measure, plot and display capacity data.
- Use scheduling to find and ease constraints.
- Reduce the number of appointment types to reduce complexity and carve-out.
- Work differently – flexible hours, weekends, pre-plan and cover annual leave, extended roles, etc.
- Bid for resources only when the constraint is equipment or staff and working differently will not help.

Other useful tools and techniques

- **Process templates**
- **Process mapping**
- **Theory of constraints**
- **Root cause analysis using five whys**
- **Statistical process control (SPC)**
- **Managing variation**
- **Discharge planning**

Background

While thinking about capacity and demand is relevant to many types of process, a number of the most useful approaches for healthcare originate from the **theory of constraints**.