NHS

Online library of Quality, Service Improvement and Redesign tools

Demand and capacity – a comprehensive guide

NHS England and NHS Improvement

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.

Analysing the demand and capacity within a service will enable improvements to be made that smooths the flow of service users through the system and helps to create a better patient and staff experience of the healthcare process. This is true regardless of whether you are looking at services providing inpatient, outpatient, community or mental health services.

When to use it

Delays within systems occur when flow into the system is greater than flow out of the system. Demand and capacity analysis allows you to identify the reasons for the delays and the part of the system that is causing the backlog. If demand and capacity is monitored on a regular basis, it provides a thorough understanding of flow through the system, and allows service leaders to identify increasing waiting list lengths, bottlenecks or constraints early on, to avoid possible future delays and the build-up of a backlog.

To do this, you need to have reliable measures for demand, capacity, activity and backlog in place. This guide will outline a process for demand and capacity modelling from the beginning, allowing you to undertake the initial analysis, after which point you can build the measures into routine monitoring and service management.

How to use it

STEP 1: Define and identify demand, capacity, activity and backlog for the service.

Some hints on how to do this can be found below.

1.1 Demand: what we should be doing

All the requests/referrals coming in from all sources, both electronic and paper-based, 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 but demand which must be met within a short time-frame. For example:

- In an intensive care unit, the capacity is limited to the number of beds available. When demand exceeds the number of beds available, something else must happen to the service user 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.

- In a community setting hidden demand can reflect care which is self-referred: if service users are phoning and are either unable to get through or asked to phone back later, rather than being put into an appointment or on a waiting list: this is hidden demand.

1.2 Capacity: what we could be doing

This refers to the resources available to do work. For example:

- In a diagnostic setting this might refer to any specialist equipment required, multiplied by the hours of staff time available to run it.
- In a community setting services might depend on a multi-disciplinary team being available, or something as simple as room availability. A particular challenge in relation to this could be ensuring that capacity is focused in the right location, for example if patient choice does not match the capacity provided, then it is possible that the backlog for a service could grow despite having unused clinic slots.
- Another particular challenge in relation to capacity setting in some community and mental health settings relates to the balance of 'same day' demand with regular treatment. In these services there may not be a waiting list or backlog but there are referrals which need to be picked up and seen either the same day or within 48 hours, alongside regular, planned work. Some organisations address this by 'weighting' each demand type/visit, and developing staffing rotas that ensure teams are safely staffed with the right skills which can flex up and down when required.

Again it is important to look at actual capacity rather than 'theoretical' capacity. Theoretical capacity can be calculated by looking at staff rotas and such, 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. Particular elements that need to be considered could be staff annual leave, mandatory and additional study leave, sickness, time out for urgent meetings, travel time between service users (particularly in community services), supervisions, management duties such as 1:1s and management paperwork/HR and processes.

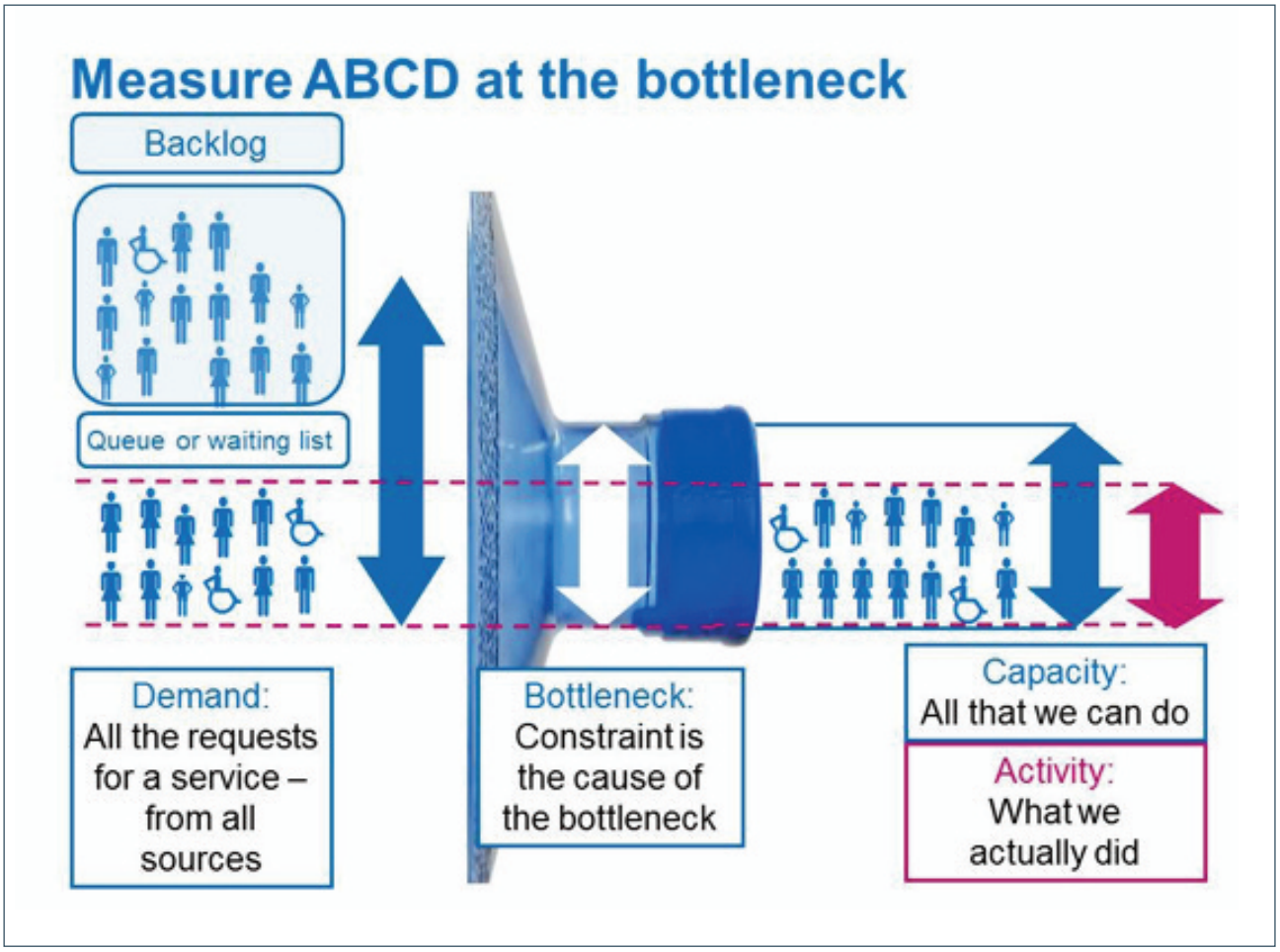
1.3 Activity: what we are actually doing

This 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: what we should have done

This is previous demand that has not yet been dealt with, showing itself as a backlog of work. A backlog is different to a waiting list, in that a backlog is unsustainable: it is the mismatch between your demand and capacity, and means that the waiting list will keep on growing, with service users waiting longer to be seen.

Figure 1: activity backlog capacity and demand



Once you have done these things you are ready to move onto the next steps.

STEP 2: Engage your stakeholders and process map

Map out the steps involved in the processes or service user pathways at a high level and then 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 mapping** tool will help you at this stage.

STEP 3. Identify the problems in your process map with your bottlenecks and true constraint

3.1 Bottleneck(s): any part of the system where service user flow is obstructed causing waits and delays.

The bottleneck is the step in the process where there are the longest delays for service user at any part of the system where service user 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 called the constraint. We often look at bottlenecks like a logjam in a river: it doesn't matter how much capacity is available after this particular step, because no further interventions can commence until this step has happened, and it is this step that is causing the delay.

3.2 Constraint(s): what restricts the capacity of the service at the bottleneck

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). Backlogs usually occur before the constraint in the process map and clear after the patient has gone past the step with the constraint. **Process templates** are a visual representation of what happens to one patient as they go through a process and help to identify the constraint allowing you to support scheduling at the bottleneck.

STEP 4: Keep asking 'why?'

In order 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. Using the five whys tool will help you to generate ideas for change that you can then act upon.

STEP 5: Measure your demand, capacity, backlog and activity in units of time

Based on the definitions described in Step 1, you can now move on to measure it, to build on your detailed process map for the patient journey you are focused on. For comparison purposes, these should all be measured in the same units for the same period of time, for example hourly, over a 24 hour period, weekly or monthly. It is also important to compare the four measures on a single graph. This is important because your team is likely to be one stage in the patient's pathway or journey. It is possible that another team's work is the bottleneck. For example:

- In order to meet the four and twelve hour targets a lot of hospitals focused effort on capacity in the accident and emergency department. However, assessment of the patient pathway showed that one limiting factor was actually the completion of a mental health assessment and subsequent availability of a mental health inpatient bed. It was this subsequent lack of availability that slowed the 'flow' within the pathway and caused the delay, not capacity within the A&E department specifically.
- In an outpatient setting the bottleneck may be a delay in the interpretation of a diagnostic report, leading to a later outpatient appointment for treatment.

- In a community setting the constraint could be the availability of a domiciliary physiotherapy service – meaning that any intervention required after physiotherapy is delayed until this service can be provided in the patient's home.
- In a mental health setting delays could be limited due to the capacity of staff availability to undertake new patient assessments – following which they could be streamed into a counselling service or group therapy, but the patient cannot enter the waiting list for either of these until the initial assessment is complete, even if there is capacity further down the pathway.

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, multiply the number of patients referred by the time in minutes it takes to 'process' the steps for a patient. (See **process mapping** tool).

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 at the bottom of this section shows how to measure demand.

When measuring demand you must include all demand sources. For example:

- In a fracture clinic you might track orthopaedics, physiotherapy referrals, pain clinic and direct orthopaedic referrals;
- In a community mental health team you not only need to track referrals from all sources (paper and electronic), but also those for new patient assessments, routine follow-up visits, and crisis work.

This will then produce a complete picture of the true demand, and the capacity it requires.

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. You should avoid using averages 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

Calculate the amount of staff time available to run the identified clinic/s in minutes, to match that for which you have measured your demand. If a specific, limited, piece of equipment or space is required (for example in an acute setting this could be a machine such as an MRI scanner, or in a community setting this could be the physical space available – such as a room in a community clinic). Multiply the number of pieces of equipment/rooms by the time in minutes available from the people with the necessary skills to use the equipment or run the clinic. For example, two treatment machines multiplied by 480 minutes (eight hours) of session time amounts to 960 minutes (16 hours) of capacity each day, as long as the required staff are available. 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 resource	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 and unavailable 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 requires careful and rigorous analysis. Here are a few questions that may help your calculations:

1. Determine the overall supply of the service – how much capacity is available. For example minutes in outpatient clinic time?
2. Consider how supply changes over differing weeks and months (eg staff leave). Consider the peaks and troughs and whether these match the peaks and troughs in demand – understanding actual capacity as opposed to potential capacity enables you to look at ways of bringing the two closer together (eg co-ordinating consultant leave may result in fewer clinic cancellations).
3. Identify how the supply is provided. Consider for example:
   - Can the time it takes for a patient going through the process be reduced by removing steps which add no value
   - Does the service work in 'batches' with, for example, patients waiting for a specific clinic, such as in the community examples earlier in this document? It can be much more efficient if supply is deployed evenly against demand because the closer demand and supply can be matched, the better the system will run.

4. Is the service providing what is really required to meet the patient's needs?

- In an acute setting for example, the provision of radiology services for the management of DVT may be provided in a more efficient way;
- In a community setting consider the locations of clinics – as in the example earlier sometimes you may have enough capacity in terms of time, but patient choice leads to a backlog building up for a service in one location, whilst that in another location has spare slots
- For domiciliary services consider the process for reviewing home visits. Ensure that they are still required as home visits, or whether a patient's needs can be met differently as the condition changes, is managed better, or improvements in technology are deployed (for example remote monitoring). Be sure that you are factoring in travel time, and that travel is arranged in the most efficient way for staff and services.

How to measure activity

Multiply the number of patients processed through the service 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. It is a useful measure to know however, for later analysis and to understand pinch points that may have caused your backlog to build.

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

For example,100 patients on the waiting list multiplied by 20 minute treatment time each equals a 2,000-minute (33.3 hours) waiting list. On a daily basis there may be more referrals coming in than the capacity can meet, therefore creating, or adding to, a 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. For example:

- Patients requiring radiotherapy treatment can be on waiting lists for their pre-treatment, planning and simulation at the same time. Only count them in the earliest stage (planning) to avoid recounting them later in the process.

If however, you run a service which has multiple waiting lists, for example Early Intervention Psychosis, where a patient might be seen for a period of 1:1 intensive counselling sessions, as well as group work and routine follow-up appointments, these should be considered as separate services for the purpose of demand and capacity planning and mapped out and calculated accordingly.

Figure 5.1: 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

Figure 5.2: How to calculate a backlog in a mental health service

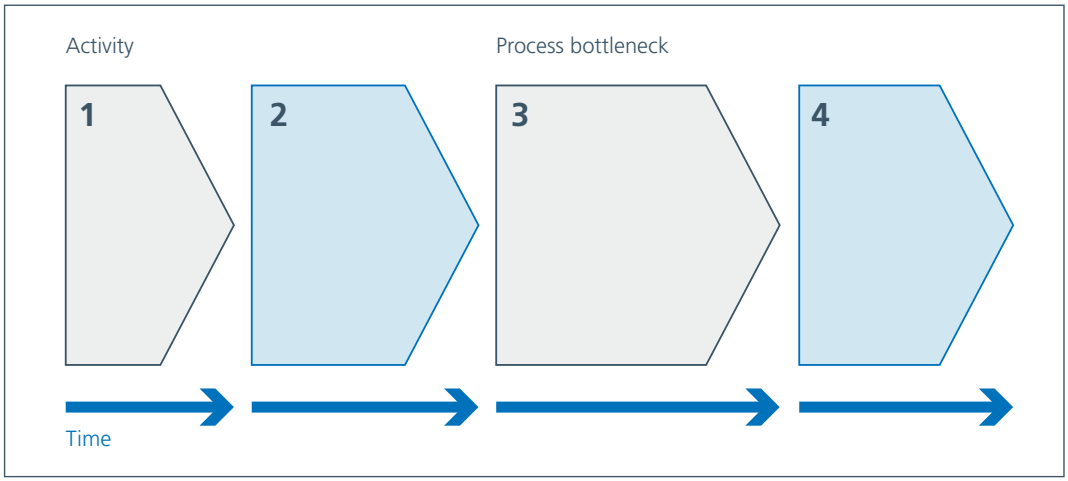
Type of request	Number waiting	Minutes taken to complete task	Minutes taken to clear backlog
Psychology assessment	6	60	360
Psychology intervention	4	60	240
Total backlog			600

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.

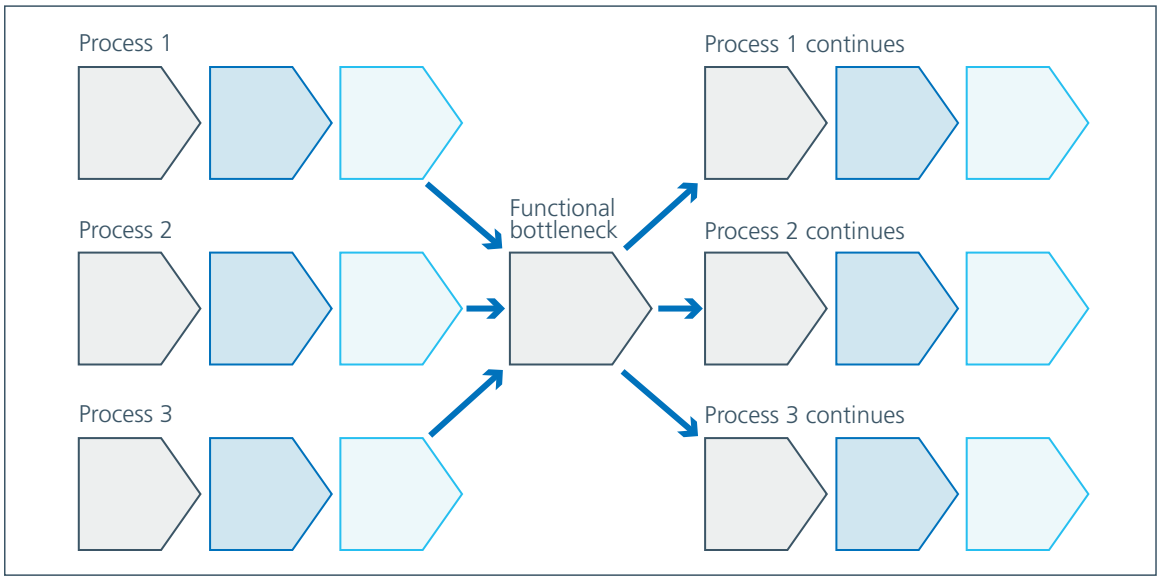
Figure 6: Process bottleneck



- **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. For example, 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, and not plan around averages. Capacity should be set at a specific percentile of demand. For more information on how to calculate percentiles please **see our video**.

The exact percentile at which you should set your capacity will depend on the urgency profile and size of demand for your service; usually it will be somewhere between the 65th and 85th centile. It may be that to identify the right percentile for your service you need to try out a few and keep reviewing them until you find the centile that is right for your service. You can use the PDSA method to do this – a guide on **how to use PDSA cycles can be found here**.

Other 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**.