Staff working in clinical departments should understand a number of key elements of the services they are responsible for. This section discusses some useful tools and concepts essential to improved process change in the management of outpatient services.

This chapter deals with these essential tools.

PRINCIPLES

- Staff managing services in health boards should have a good understanding of the capacity of their service, the activity levels provided by the service, the demand on the service, and the backlog of work in the system.

- Managers should understand how patients proceed through the service.

- Managers should understand the flow of work through the service, and be able to identify rework within the system, constraints and bottlenecks, and unnecessary process steps.

Process mapping

It is important to understand how patients proceed through the service. This is covered in more detail in section 4.2.

Analysis of activity, backlog, capacity and demand

This helps to understand how the service can cope with patient demand. This is covered in more detail in section 4.3.

Understanding follow-up demand

Understanding the demand for services is fundamental to knowing how to deal with them. This is covered in more detail in section 4.4.

Measuring follow-up demand

Follow-up patients are one of the main demands on outpatient services. Before we can reduce follow-up demand, it is important to understand it and measure it. See section 4.5.

Reducing follow-up demand

There are a number of tools to manage and reduce the demand on outpatients created by follow-up appointments. This is covered in detail in section 4.6.
Constraints and bottlenecks
It is necessary to understand two other concepts in order to manage capacity and demand: constraints and bottlenecks. This is covered in section 4.7.

Flow Models
Flow models are a way of analysing the service constraints and bottlenecks found during the mapping process, and looking at them from the perspective of the service; they also help in determining whether work is scheduled around the constraint. Further details can be found in section 4.8.

Carve out – understanding queues
Patient experience is often about waiting and has been described as long periods of waiting, punctuated by short periods of activity. Understanding the principles behind queue management will reduce queues and improve waiting times and ultimately the patient experience. This is described in more detail in section 4.9.
Process mapping is ubiquitous in the NHS improvement movement. Understanding the process from the patient perspective is essential if services are to be improved.

There are two stages to process mapping. First, understand what happens to the patient, where it happens and who is involved. Then examine the process map to determine where there are problems such as multiple hand-offs, parts of the process that are unnecessary or do not add value, or parts of the process which could flow better if undertaken in a different order. These problems can be addressed by designing a new more streamlined process.

Second, use process mapping to determine where bottlenecks and constraints occur. Is use of the constraint maximised? Do the patients flow through the system without delays? This approach is covered in Chapter 4.8.

Patient processes in healthcare
Patient processes have evolved over time within the NHS with many new systems being bolted on to current processes without an overall analysis of how the whole system functions. There can be many layers to a patient’s journey and no one professional has a detailed overview of that journey. It is therefore essential to any improvement work that current systems are mapped by the staff on the ground floor who are directly working within those systems, and with the involvement of patients and carers who experience the whole journey from a very different perspective.

Benefits of process mapping
Process mapping is a simple exercise, and ensures any improvement work is based on a realistic analysis of current working systems, as opposed to how local policies determine they should be working. It provides a clear indication of where there may be impact on other parts of the service when changes are made, as well as the opportunity to get multi-disciplinary teams from across the healthcare community together to promote a culture of ownership and continuous improvement. Staff are often not aware of the complexity of a patient’s journey and this provides an opportunity for staff to understand how their work impacts on other parts of the system.

The final map itself can be used as a training tool; for communication purposes as well as identifying areas for continuous improvement. The map should be updated when changes are made to ensure current processes are being captured.

The high level process map
When choosing which patient groups to map, the group should be of high volume and share common characteristics. This provides the opportunity to have the greatest impact. This stage of the process map can be used to standardise care or treatment. (see figure 4.1 on page 48)
Choose clearly defined start and end points; for a referral process these may be the arrival of a referral letter in the health board through to the appearance of the patient in the outpatient clinic. The important thing is to be clear which parts of the process are within the scope of the map and which are not.

At this stage a quick mapping exercise by a few staff may be useful to determine who will be involved in a more detailed mapping exercise. It is essential to have representatives of all staffing groups involved in the process at the main mapping exercise, and a quick high level map will help ensure no staff group is forgotten. Do not forget to involve patients in the detailed mapping process.

Organising the mapping event
Once you have a high level process map of the journey you wish to analyse, check you have a list of all staff groups that are involved in the patient’s journey. This will ensure that further down the improvement process when changes are to be implemented, you will already have the staff groups that are involved in the process on-board and will not have to spend time engaging staff who have not been involved from the start.

Mapping workshops should be at least a half-day event with all members of staff invited for the whole session, although a drop-in session can be useful where it is impossible to get all staff groups together. In this situation it is vital that all staff involved have an opportunity to validate the final map. If you can get all the staff together there may be time for some action planning at the end of the session, to agree the next steps to the improvement process and how they should be involved.

It is advisable to use an independent facilitator to run the event, and it may be useful to run the session off-site, to facilitate neutrality. At the outset of the session, ground rules should be set with the group, for example:

- Freedom for everyone to be open and honest.
- Mapping reality versus what should be happening.
- Focus on what usually happens for most patients.
- Don’t make assumptions about people’s knowledge.
- Clarity over what is being mapped.
5 minute ruling — if there is debate over a certain step, note the issues and move on after a 5 minute discussion.

The session is looking at processes not people.

Throughout the session it may be useful to reiterate some of these rules, and it is essential that there is a blame free culture present. It is a human response to be defensive when groups are looking at the way people work, and the group needs to ensure all staff are focusing on the processes and not the staff involved in those processes.

At the mapping workshop, use ‘post-it’ notes to capture the information about the patient journey down to the level of one person, one place and one time. See figure 4.2.

This will ensure that hand-offs, multiple staff, changes in location, and loops in the process are all captured. Capture any issues or suggestions for improvement that are made on a separate flip-chart to circulate with the map when finalised for comments. Arrange the ‘post-its’ into order, and look for:

- Things that are done more than once.
- Steps that do not add to the patient outcome — ask “Why is this being done?”.
- Count the number of hand-offs.
- Identify where there are delays, queues, and waiting built into the process.
- For each step ask whether the action is being undertaken by the most appropriate staff member.
- Look for ‘re-work loops’ where activities are taken to correct situations that could be avoided.
It may be useful to re-draw the process map to look at a specific issue. For example, a process map can be drawn with each staff group in a different row or column to identify the hand-offs—a hand-off occurs each time the process map moves across to a different column. (see figure 4.3 below left)

**Focusing in on the problem**

Once the overall process map has been drawn and the staff agree with the process, it is useful to identify where there are bottlenecks, constraints, and steps that do not add value in the process.

Which step causes the most delays? This step can then be mapped in more detail, expanding out the process. This can be done several times, each time expanding and getting to a greater level of detail. (see figure 4.4 right)

Any level of mapping is useful, and it is rare for a group to undertake process mapping without identifying at least one step that some members of staff were unaware of. The best way to learn process mapping is to do it.

Some simple tips:

- Try photographing key steps of the process and illustrating the map for a staff presentation.
- Walk through the process with a patient to check that all steps/events are included.
- Only consider what usually happens.
- Involve all staff groups and patients.
Next steps:
Process mapping is the first stage, helping to identify where to start making improvements. The use of Plan Do Study Act (PDSA) cycles of improvement provides a structured approach and framework for developing, testing and implementing changes. View more information on process mapping.

When identifying potential areas for improvement look for the following:

- Reduce the number of hand-offs and steps within the process.
- Co-ordinate the scheduling of appointments for patients with complex needs.
- Create trigger systems so that booking diagnostic tests trigger an appointment for results where appropriate.
- Reduce the number of times a patient has to attend hospital.
- Reduce or eliminate batching; this often creates delay in a process.
- Reduce the number of queues to be managed; see section 4.9 for more information.
- Extend staff roles to encourage flexibility.
An understanding of the dynamics of waiting lists is essential to managing them. There are four key measures that must be understood and monitored on a continuous basis if waiting lists are to be managed effectively. It is also important to understand the two key types of limitations in the system; constraints and bottlenecks.

**Backlog**
The backlog for outpatients is the number of patients on the waiting list.

**Capacity**
The capacity of the system is the time that the resource is available. For outpatients, capacity is staffed clinic sessions.

**Demand**
The demand on the service is all the patients referred into the service from all sources.

**Activity**
Activity is the throughput of the system – the number of patients seen in clinic.

Consider the following scenario: waiting lists are increasing, and it is necessary that they are reduced. The demand is 5000 patients per annum, and only 4500 patients are seen in clinic. Waiting lists are going up by 500 patients per annum. What should be done? The health board is asked to fund an additional 500 cases.

Consider another scenario: there is an increase in waiting times for CT scans. The wait has gone from 12 weeks to 18 weeks over the past two years. The data supports the impression that there are more referrals. What should be done? The health board asks for an additional CT scanner. There is an unstated assumption behind both of these scenarios. The consultant in clinic works 100% of the time. The CT scanner is utilised 100% of the time.

Capacity is the ability to do work, not the amount of work done. It may be true that the CT scanner is working at 100% capacity, but without data for both activity and capacity, that assumption cannot be made.
The NHS collects data on activity, but rarely on capacity. To understand capacity, we need to understand the relationship between activity and capacity. (see figure 4.5 below)

Capacity and activity should be measured in a common unit of measurement to understand how they relate to each other. Time is often used.

**Capacity**
Capacity is the resource available, multiplied by the time it is available. In practical terms, outpatient capacity is a combination of physical capacity, staffed capacity and clinician capacity.

**Activity**
Activity is the actual work carried out and usually measured in patient numbers. However, it is not possible to compare two items measured in different units. So if the intent is to compare activity to capacity, both activity and capacity should be measured in time. In the case of outpatient activity, this is the time it takes to see each patient (usually longer for new patients than follow-ups) and the number of new and follow-up patients attending, multiplied by those times. Activity is measured as the total number of patients processed, multiplied by the time taken to process each patient.

---

**Figure 4.5**
The four key measures

![Figure 4.5](image_url)
**Demand**

In simple terms, demand is the volume of referrals. However, this doesn’t take into account patients removed from the waiting list other than for treatment. Derived demand resolves this issue as it is calculated using the following formula: Waiting list on date B – Waiting list on date A + activity occurring between date A and B.

**Backlog**

Backlog is the previous demand that has not yet been dealt with, showing itself as a waiting list or backlog of work or waiting list.

**So why do waiting lists and waiting lists or backlog of works form?**

The main reasons why waiting lists or backlog of works develop is the mismatch between variation in demand and capacity at specific times, because the right people or equipment are not always available to deal with the demand in a timely manner. Every time the demand exceeds the capacity, a waiting list or backlog of work is formed.

The waiting list is also measured in terms of time. Estimated times for each treatment/appointment can usually be obtained. The appropriate time can then be allocated to each patient on the waiting list, and the total waiting list expressed in appointment minutes can be captured at the end of each month. Be careful not to use the average appointment time.

**The common graph**

Figure 4.6 shows the four measures for one such list, plotted on the same graph. The relationship between the four measures can now be seen. However, including the backlog can hide the detail of the other measures – if the activity, capacity and demand lines are very close together, consider removing the backlog. In this graph capacity remains constant each month. In reality there may be changes in capacity depending on the availability of staff, clinic rooms and other resources.
Waiting lists and waiting times build up because of variation in demand and capacity, (where demand temporarily exceeds capacity). The mismatch or variation in capacity and demand is one of the main reasons why waiting lists and waiting times increase.

Once demand, capacity, activity and backlog have been measured the data and patterns that emerge can be used to start predicting demand and managing capacity, activity and backlog at the bottleneck.

Looking at, and matching capacity and demand is a key approach to removing some of the visible and hidden backlogs along the patient pathway.

Variation is a normal part of all processes and clinical processes are no different. Accounting for variation is important when identifying the throughput (total activity) of a service and matching the capacity to manage the demand. Averages hide some of the variation. When calculating throughput do not use averages; using average appointment time to calculate the waiting list will usually underestimate actual demand.

Rather than using the arithmetic average or mean (50th percentile) you could use the 80th percentile.

To calculate this, for example take 100 patients. The 80th percentile can be found by sorting the patients by the length of their appointment then counting to the 80th patient. There are software packages that will calculate percentile.

By using the 80th percentile, you are likely to have underestimated the time needed on only 20% of occasions. It is far less likely that you will have a run of patients over the estimate. Using the arithmetic mean or average length of appointment, will generally be shorter than this and therefore could be more likely to underestimate the time needed. See Figure 4.7.
Understanding the demand for services and how it balances against the activity the system is producing, is fundamental in understanding where waiting lists come from and how to deal with them. This section addresses issues of activity levels, changing demand and how to relate one to the other.

Seasonal variation
Are referral rates stable across the year? In some clinics, referrals increase at certain times — typically medicine referrals increase in winter, as do those for orthopaedics. Referrals in dermatology increase in summer. Some specialities such as ophthalmology do not have obvious trends. Seasonal trends are important as they can skew the analysis — if you do not look at the whole year, is the increase you have detected a real increase, or is it due to a summer bulge? When looking at long term trends it is useful to compare years on the same graph, so that changes between months can be separated from the year to year changes. An example can be found in figure 4.8 where waiting lists for several years are compared. Note that the newest line is stable — compared to an upward slope for the same months of previous years.

Conversion ratios
It is important to treat the elective process as an integrated whole. Increased activity at outpatient clinics may affect the flow on to the surgical waiting list, but simple conversion ratios may not give accurate predictions of future surgical demand.

With long waiting lists, an initiative clinic to remove patients from the tail of the waiting list where PTLS has not previously been used, may deliver lower conversion ratios due to the nature of the patients who tend to accumulate at the end of waiting lists. Conversely, pre-screening of orthopaedic referrals by a physiotherapist may considerably reduce the number of new referrals onto the consultant outpatient waiting list. However, if the consultant continues to see the same number of outpatients, the ratio requiring surgery and thus the number flowing onto the inpatient waiting list, will increase substantially as those patients unlikely to proceed to surgery have been screened out.
Many health boards and specialities have a problem with seeing outpatients at short notice. In response to the need to fit patients into full clinics two or less weeks into the future, health boards have traditionally set aside appointment slots, or overbooked clinics at the last minute.

Understanding the workload.
Clinics are often overbooked and capacity is often exceeded by demand, leading to overcrowding and patients not being seen within the timeframe that is clinically appropriate. This is often referred to as ‘follow up not booked’. Figure 4.10 shows that there are no available slots for the next seven weeks! Even where information about the number of empty appointment slots is available when booking, one vital piece of information is missed out: how many more patients will be added to that clinic between now and the clinic taking place?

One way to understand what the demand will be on future services is to prospectively record every follow-up appointment made for a time period (preferably several months) and keep track of how many one week, two week, three week etc. appointments are made each week. An alternative is to estimate the demand based on historical data. By analysing all the appointments made over an 18 month period, and calculating the time between the appointment and its predecessor, an estimate of appointment frequency can be made. The distribution for one ENT service is shown in figure 4.11.

Note that 5% of appointments are made for one week, 5% for two weeks, 5% for three weeks etc. 25% of appointments are made for less than six weeks. In order to leave room for these patients, at six weeks the clinic should be only 75% full.

Figure 4.10
A typical ENT clinic

Figure 4.11
(far right) Number of weeks/months between follow up appointment being made and the appointment
The clinic in figure 4.10 (see page 57) is already full for the next seven weeks. Where will these patients be placed?

There needs to be a way to look at clinics and take account of the work that will come in, not only the work that has already come in. If this information is made available to clinicians, then they will be better able to make decisions about when to bring patients back — balancing clinical need with clinic availability, helping to avoid ‘follow up not booked’.

**Step 1: Plot the distribution...**

As long a time frame as possible should be used for this exercise; the shorter the time you choose, the more under-represented the longer appointments will be, as either the first or the second appointment will be missing from the sample. Use a sample size of 18 months; more would be better! See figure 4.12.

**Step 2: Convert to a cumulative graph...**

The lower area represents the proportion of the clinic that should be filled, based on the distribution in the first graph. The upper area is the proportion needed to deal with the appointments ‘in the future’. The boundary between the upper and the lower is the proportion of the clinic that should be filled at any time. The boundary line represents ‘full’ if there is to be space available for any appointments in the future. Figure 4.13
Step 3: Plot your ‘actuals’...
The jagged line represents the ENT clinic seen overleaf. As can be seen, there are times when it is above the boundary line — the clinic is overfull. There are also times when the clinic is underfull for that week or month. Figure 4.14

Step 4: Simplify the presentation...
The last graph ‘flattens’ the boundary line, setting it as ‘100%’; see Figure 4.15. The jagged line becomes the series of vertical bars, showing over and underbooking against the new profile. This graph could be given to consultants or clinic staff to show where problems are predicted.

Understanding demand for follow-up appointments will assist booking, and it will go a long way to improving the current chronic overbooking found in many clinics.
Proactive management of follow up appointments is as important as new appointments, both in terms of volume and management of clinical risk. Every follow-up appointment uses valuable resource and should be avoided where clinically appropriate. This will help to reduce overall demand. This in turn could facilitate improved management of clinical risk by enabling patients to be seen within the timeframe that is clinically appropriate.

There are many ways to reduce follow-up demand, this in turn will improve productivity of appointment slots and create capacity for ‘new’ appointments which should reduce waiting times.

**PRINCIPLES**

- The use of patient initiated follow up appointments for individuals with chronic disease or recurring problems facilitates the patient having control over their health and well being.

- Understanding each clinic profile is an essential part of managing the service.

- An understanding of frequent attenders helps to facilitate improvements in managing these individuals.

- All departments should develop follow-up protocols for their major patient groups, and monitor their use.

- An outcome should be recorded within the Patient Activity System (PAS) for every patient interaction.

- Implementation of clinical pathways that pro-actively avoid follow-up.

**Understanding your clinic**

It can be difficult to see the wood for the trees. Sometimes it is necessary to step back and review what is happening in outpatient clinics as a concerted exercise, rather than relying on perception and anecdotal evidence.

One way of doing this is to create a ‘Week in the life of a clinic’ to determine just what happens during a typical week.

The basic technique is to set up a data collection tool that can be applied to every patient seen in a department in a typical week. The data can be collected from the PAS, from patient records, and from checklists completed by staff during the clinic. Each of these will give a part of the picture:

**From the PAS**

Patient demographics including age, gender, postcode analysis, referrer analysis.

**From the patient record**

Diagnosis and co-morbidities, number of times the patient has attended in the last year.

**From a checklist or electronic outpatient activity system**

Who saw the patient, action taken, tests ordered, when the next appointment will be.
The information collected from these sources over a typical week should give enough data to be significant. Combining the data into a single database will allow collation across the different sources. The aim of the collation exercise should be to provide a picture of a typical cross section of patients seen in the department, and should help in answering a number of questions such as: what proportion of patients are local? What is the level of co-morbidity? What disease groups make up the highest proportion of frequent attenders? How many of the follow-up appointments were seen by junior staff, and what were the decisions taken?

The analysis may not provide any quick answers, but it will contribute to any subsequent improvement process. Without knowing the nature of the thing that is to be improved, there is a danger of concentrating on what seems obvious, or what ‘everyone knows is the problem’. Having good data on the current situation is the first step to improvement.

This exercise can take a significant resource. Collection will involve looking at patient notes, and staff completing checklists for every patient they see during the week. The benefits are well worth the effort.

**Frequent attender analysis**

From at least a year’s data download from the PAS, do a frequency analysis of attendances by patient. The only data required for this is the patient number and the date of the outpatient appointment. Within a spreadsheet, a pivot table can be used to count the number of attendances by each patient. The pivot table can then be used to count the number of patients who had one appointment, two appointments and so on.
This is likely to show that the number of patients with significant multiple appointments is low, but that the number of appointments taken up by those patients is significant. In figure 4.16 (page 61), which represents data for two years, one patient had 28 appointments, and three had 24 or 25 (one every month). In all, 18 patients (0.20%) accounted for 343 appointments (1.81%), below line ‘A’ while 125 patients (1.40%) had 10 or more appointments over the two years, representing 1,615 appointments (8.53%) below line ‘B’. The graph in figure 4.17 shows the same data.

What can be done with this data? Pulling the patient records of the 18 patients and doing a clinical review may reveal changes to the care which would result in fewer appointments. Are these patients best managed in outpatients? Would there be benefit in meeting with the patient’s GP? By focusing on patients who are frequent attenders it may be possible to customise their care and improve the outcome, while also reducing the demand on outpatients. The key is identifying the small number of patients who are consuming relatively large amounts of clinic time, and then seeing if they can be managed in a different way.

Use of patient initiated appointments
Many departments now use self referral for follow-up rather than fixed appointments. This can be effective in reducing both follow-up appointments and DNA rates for some patient groups. ‘Seen on symptom’ (SOS) appointments can be used for patients with a recurring problem, where a procedure is being followed up, or where monitoring of a chronic condition can be undertaken in part by the patient.

Recurrent problems
This is the most common use of SOS appointments. There is little point in regular follow-up of a condition which flares up from time to time, which does not usually coincide with an appointment. It is preferable to have rapid access to an appointment when the problem recurs.
**Procedure follow-up**
In many situations it is common to recall patients while they are recovering from a procedure. Often appointments are made on a regular basis, until the patient reports no further problems. This will always result in one more appointment than necessary (when the patient attends to report no symptoms) and often more than one appointment.

Informing the patient about the progression of recovery (through a patient pathway), and allowing the patient to make an appointment quickly if there is a deviation from the pathway or if the patient has any concerns, will mean that patients on the normal recovery path will not take up clinic time, which can then be devoted to those patients with problems.

A value based approach to healthcare that uses resource wisely, avoids wasteful follow-up appointments will help to reduce waiting times.

**Chronic condition monitoring**
Patients with chronic conditions can also use SOS appointments, with the possibility of less frequent scheduled review. This type of appointment helps to give the patient control over their own health and well being.

In Aneurin Bevan University Health Board patients with multiple sclerosis were given the opportunity to use SOS appointments to help manage their condition. This reduced the follow-up waiting list from 333 to 92 over a three year period and the number of follow-up waiting beyond target date reduced from 249 to 16. Read the case study.

In the case of chronic conditions, as with procedure follow-up, it is essential that the patient understands what events should trigger an SOS appointment and instructions on how to make an appointment.

**Follow-up referral protocols**
Agreement amongst clinicians and the utilisation of protocols for follow-up appointments will help to avoid unnecessary follow-up appointments.

In Aneurin Bevan University Health Board patients with rheumatoid arthritis are managed jointly between GPs and secondary care – reducing the number of follow-up appointments required in hospital and allowing patients to be seen closer to home by their GP. Prior to a 3 year transformation programme, routine patients were waiting beyond 28 weeks, contravening NICE Quality standards. Now nearly 85% patients with suspected inflammatory arthritis are seen and treated within 6 weeks of a GP Referral. Read the case study.

**Attendance outcomes**
An outcome should be recorded within the PAS for every patient interaction, whether the patient is present or not. This should help to reduce the need for administrative validation, and appointments that are not required.
Two other concepts help to manage capacity and demand: constraints and bottlenecks.

**Constraints**
The constraint in the system is the factor that ultimately restricts the capacity of the system, i.e. the rate at which patients go through the system. In outpatients, the constraint may be physical space, equipment or skilled staff to provide appointments or procedures. The constraint is not easily removed without substantial investment in terms of staffing, or facilities.

Once identified, the constraint should become the most important part of the process. Work should be scheduled so that the maximum use is made of the constraint. Resources at the constraint should not be used for jobs that other staff/resource could do.

**Bottlenecks**
A bottleneck is the part of the system that restricts the activity. A backlog tends to happen just before a bottleneck and clear after the patient has gone past that stage.

A typical bottleneck in outpatients may be patients booking in at reception. The entire system stops while waiting for a patient to move from the reception because of a shortage of reception staff or queuing to use a technical solution for booking in.

Distinguishing between the constraint in the system and the bottleneck currently limiting activity is essential. **Constraints limit capacity (resource available) and bottlenecks limit activity (number of patients processed).** By removing bottlenecks it is possible to increase activity until it gets close to the capacity of the system.

**Constraints and bottlenecks: a three step process**

**Step 1.** Identify the constraint in the system. Use process mapping to determine where the constraints are.

**Step 2.** Determine whether the process is scheduled around the constraint. Use patient flow modelling to determine this.

**Step 3.** If not, use PDSA cycles to eliminate a bottleneck or redesign the system to maximise the constraint. Then repeat Step 2.

When the point is reached where the use of the constraint is maximised, capacity should be analysed to determine whether it is sufficient. If it is not, then it is time to consider resources, but now the data is available to support the business case.
Managing patient flow 4.8

Flow models look at the care process from a clinical unit perspective, bringing together a number of patient process maps to look at work flows through the unit.

**Principles**
Flow models are the best way to analyse the work of a unit, such as an outpatient clinic. The process of building up the model is simple if all the steps are followed.

1. **Map and agree the process**
The process must be mapped to a high level of detail.

2. **Time the steps**
For a session, record the times for each step of the process.

See figure 4.18 The flow data

### Figure 4.18
The flow data

<table>
<thead>
<tr>
<th>Patient</th>
<th>booking in time</th>
<th>start of pre-appointment checks</th>
<th>end of pre-appointment checks</th>
<th>start of appointment</th>
<th>end of appointment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>08:30</td>
<td>08:45</td>
<td>09:01</td>
<td>09:23</td>
<td>09:30</td>
</tr>
<tr>
<td>2</td>
<td>08:33</td>
<td>08:03</td>
<td>09:13</td>
<td>09:36</td>
<td>09:46</td>
</tr>
<tr>
<td>3</td>
<td>08:44</td>
<td>09:15</td>
<td>09:31</td>
<td>09:50</td>
<td>10:05</td>
</tr>
<tr>
<td>4</td>
<td>08:57</td>
<td>09:40</td>
<td>09:56</td>
<td>10:11</td>
<td>10:15</td>
</tr>
<tr>
<td>5</td>
<td>09:06</td>
<td>10:05</td>
<td>10:01</td>
<td>10:20</td>
<td>10:31</td>
</tr>
<tr>
<td>6</td>
<td>09:15</td>
<td>10:30</td>
<td>10:13</td>
<td>10:38</td>
<td>10:52</td>
</tr>
</tbody>
</table>

3. **Identify the constraint**
The constraint is that part of the process which is the ultimate restriction on the amount of work that can be done.

4. **Draw the flow model**
Using graph paper, or a spreadsheet programme, draw each patient as a horizontal bar one above the other. Set the horizontal axis to represent time, with the start of the session on the left, and the end on the right. Colour each stage of the process a different colour — it may help to colour the constraining process red. The length of each line will now represent the time each step of the process takes, and multiple patients will show as a series of horizontal lines. Add up the total of the red sections, and calculate it as a proportion of the total time.
5. Ask Why...
Over the course of the outpatient day, appointments took up 38% of the available time (actually even less, because the session ended 18 minutes early). This is not a good utilisation of a very expensive resource. Why are there long periods when there is no clinic taking place? What else is happening during this time?

6. Ask How...
How can the situation be improved? It should be possible to schedule the clinic in such a way so as to maximise the use of the scarce resource, which is a clinician seeing a patient. What is currently being done in clinic that could be done elsewhere? Can tasks be undertaken in parallel?

7. What is possible?
It is possible to put together an ‘ideal’ flow model scheduling around the constraint. Use estimated time for each stage based on the 80th percentile time for each step of the process. This will make it possible to estimate the optimum use of the constraint and give a target to the improvement project. Figure 4.19 the ideal flow model.

Conclusion
This analysis tool will not show what should be done, or even what the causes of the delays are. Process mapping will provide that information. The tool will show how well scarce resources are used, and how much room there is for improvement.
‘Carve out’ is an insidious process that steals capacity before our eyes, while appearing to protect the capacity for those patients who need it. In complex processes like health, some carve out is inevitable, but it must be eliminated where it can be, and managed where it cannot.

What is carve out?
Carve out is a term given to circumstances where reserving some of a resource for one group reduces the resource available to another group. Carve out is seen every day. It is present in supermarket car parks (parent with child parking), in the supermarket itself (basket only queues), on the road (bus lanes) and in health. Health is the natural home of carve out. The NHS has adopted it as a solution to a problem, and in doing so created an even bigger problem.

Carve out should be reserved for urgent patients and subspecialisation.

How many queues are there?
In a typical outpatient clinic there are many queues. There are slots for new patients; there are slots for urgent new patients. There are post-operation slots and there are soon review slots. In some clinics the number of slots can run well into two digits. These slots are created in an attempt to balance out the capacity to match the patients coming through the clinic.

The odds that every week (or any week for that matter) the exact proportion of patients will match the available slots are minute. What happens instead is that there are empty slots, and then the clinic is overbooked to fit the extra patients in. The schedule goes out of the window, and the flow of work is totally disrupted.

What is the impact of carve-out?
Figure 4.20 represents an outpatient clinic with a high degree of carve out; separate slots for each patient type, different clinics for different conditions. The lower, red line shows what the waiting list would have been if each patient was booked into the next available slot instead of the allocated speciality slot.
Carve out: understanding queues 4.9

Dealing with carve out
Some carve out is necessary and has benefits in spite of the negative impact on waiting times. Two such examples are carve out to reserve space for urgent patients in partial booking, and carve out to allow for essential clinical subspecialisation. The important thing is to allow the benefits and manage the carve out to minimise its effects.

Dealing with the 10 day waits
Partial booking allocates patients to clinics about four weeks before the appointment. There is a need to reserve some capacity for those patients that the health board does not know about four weeks from the date of the clinic.

There are ways of managing this carve out so that it does not affect the waiting times for non-10 day patients. The key to resolving carve out in this case is to manage the impact of carve out.

Managing subspecialisation
Chapter 2 provides a methodology that allows health boards to deal with carve out caused by subspecialisation. Subspecialisation has benefits in improving skill mix in specialist areas. Carve out caused by subspecialisation must be managed, rather than trying to prevent it.

Prioritisation and carve out
It is important to distinguish between clinical prioritisation and carve out. While there are issues with the type of prioritisation used, if waiting lists are longer than a few weeks, some degree of prioritisation will be essential. The degree of prioritisation should be minimised, so that as few categories are used as are required to meet the need to see patients within clinical priority.

Accepting clinical prioritisation does not mean allocating carved out slots to each category of prioritisation. This is the key to managing carve out caused by prioritisation; prioritise a single list of patients, and then allocate those patients from the top of the list into non-differentiated slots in the clinic. ‘Urgent’ patients do not go into ‘urgent’ slots; and ‘routine’ patients are not booked into ‘routine’ slots. Instead all ‘urgent’, and ‘routine’ patients are booked into generic outpatient slots. The ‘urgent suspected cancer’ (USC) patients are allocated to the 2 week slots; the volume required should be based on previous demand and should be regularly reviewed to ensure that the number of slots matches the required demand. Figure 4.21