Why do we get Queues and Waiting Lists?

A basic introduction and training guide.

To be used with the presentation ‘why do we get queues and waiting lists.ppt’

A guide to using the models

www.steyn.org.uk/models/demand analysis.xls

www.steyn.org.uk/models/demand analysis2.xls

Authors:
Kate Silvester BSc MBA FRCOphth
Richard Steyn MS FRCSEd(C-Th) FIMCRCSEd MRCGP

With grateful thanks to all our colleagues in the UK National Health Service who continue to work with us and supplied the examples.

April 2008

Why do we get queues and waiting lists?

Issue:

Queues (lines) and waiting lists are a perennial issue for healthcare organisations and for the UK National Health Service (NHS) in particular. Patients expect and are expected to wait weeks for appointments and procedures at every step in their healthcare process.

Background

Much progress has been made by setting targets for waiting times in the planned care System. In the UK, patients should be offered an appointment within 48 hours of telephoning their general practitioner and by December 2008 they will not wait longer than 18 weeks from the date of referral by their GP to receiving the first definitive treatment by a specialist.

However most NHS organisations aim to meet this target by trying to prevent patients accessing care or 'forcing' more patients' through the System to eliminate the current backlog or waiting list.

This guide is for those who want to understanding why the waiting lists occur in the first place and preventing them from happening again.

Training guide

This training guide, Mr Richard Steyns' models and accompanying the presentation (Why do we get queues and waiting lists.pps) can be found at www.steyn.org.uk.

These training materials are for those interested in understanding why queues and waiting lists occur and the principles for preventing them. It is not meant as a comprehensive guide, but as an introduction and training tool.

NB. The models should not be used to model real life scenarios with real life data. They are for illustration and training purposes only.

Why?
The statistical distribution used in these models is a flat distribution i.e. equal probability for all numbers. In the real world, the distribution for demand and capacity in healthcare is not flat, rarely normal and often skewed. Therefore is these models are used with real life data they will give a distorted picture of the potential outcomes.

How to use this guide:

This guide is to be used in conjunction models to be found at www.steyn.org.uk.

You will need to follow the instructions below or on the web site to be able to run the models on your computer.

Got to the web site and open the two models:
Demand analysis.xls
Demand analysis1.xls

For each model you will need to load the Analysis Toolpak into Excel to make the models work.

Go to:
Tools
Add Ins
And check (tick) Analysis Toolpak
& check (tick) Analysis Toolpak VBA

Click OK.

Then go to
Tools
Options
Calculation
Check (tick) Manual
And change the Iterations Minimum to '2' and Maximum to '2'
Click OK

You may then have to close the model and Save the changes and open the models again to get them to work.

Remember when you use the models on your computer the programme random number generator in your computer will produce different numbers from ours. The result will be your graphs will looks similar to but not the same as the ones in this document.

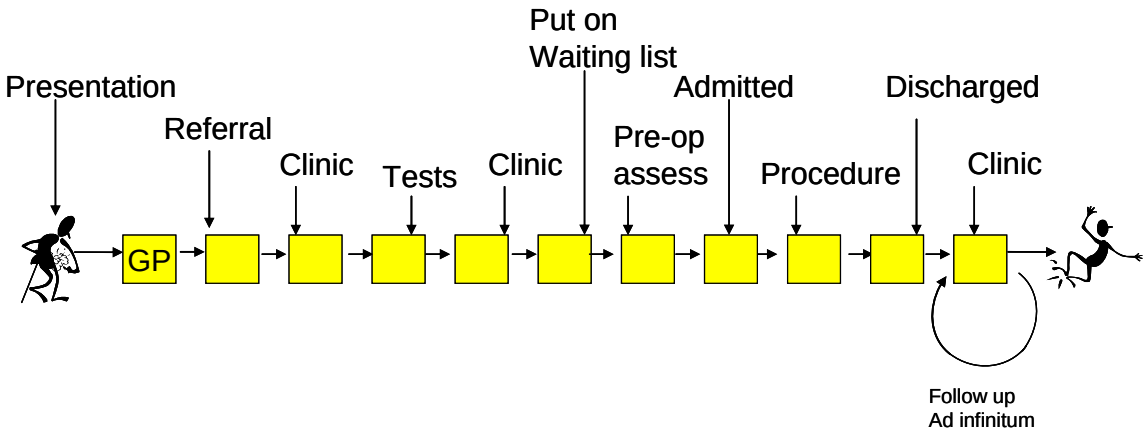
NB. The models should **not** be used to model real life scenarios with real life data. They are for illustration and training purposes only.

The statistical distribution used in these models is a flat distribution i.e. equal probability for all numbers. In the real world, the distribution for demand and capacity in healthcare is not flat, rarely normal and often skewed. Therefore is these models are used with real life data they will give a distorted picture of the potential outcomes.

The basics of demand and capacity.

Let us consider the healthcare process as a series of steps. Each step is an action performed by one person, in one place, at one time, to the patient and/or their information. Patients or their information will wait at every step.

Process map



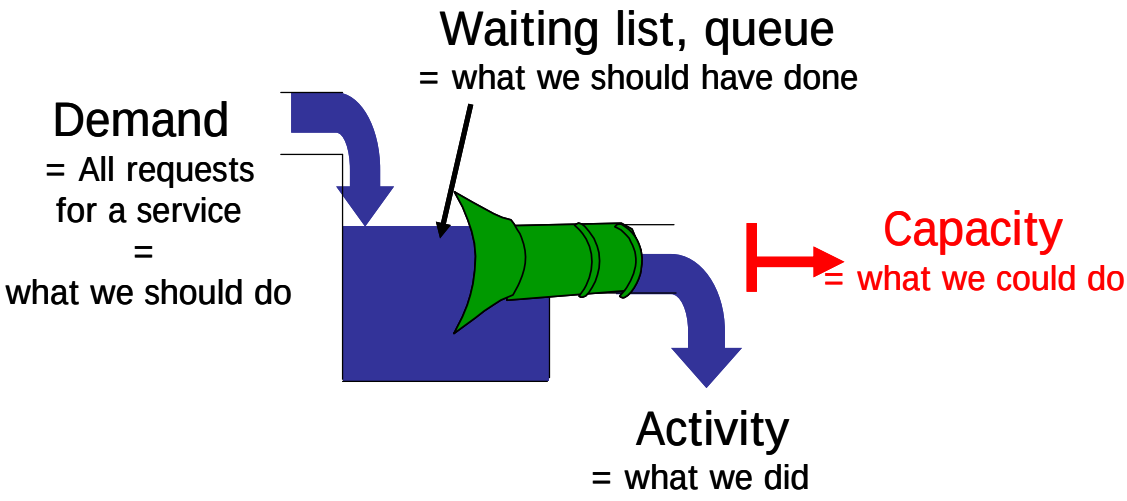
At each step there will be:

- a demand – the number of requests
- a waiting list, queue or backlog of patients who are waiting
- The activity which is the number of patients who have been processed at the step
- The capacity of each step is the number of patients that it is possible to process

Why do we get queues and waiting lists?

The diagram below shows the flow through one step in a process. The demand is all the requests for this step, the backlog is the queue or the waiting list, the capacity is what the step could process and the activity is what did go through. A bottleneck is holding back the flow.

Demand, capacity, backlog (queue, waiting list) and activity at one step in a process



To explain how these bottlenecks behave and why we get queues at each step, Richard Steyn has developed some models that can be used in Excel.
www.steyn.org.uk/models

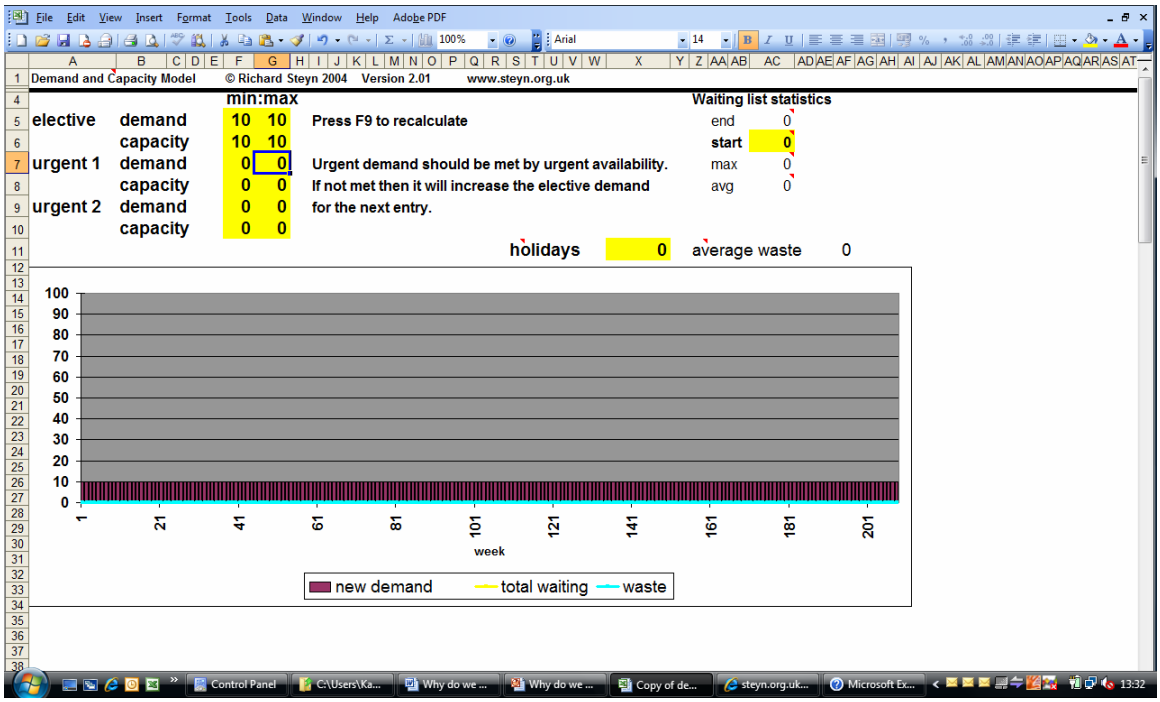
Model 1:

www.steyn.org.uk/models/demand analysis.xls

In this model we are going to consider what happens at just on step in the process. For example, a clinic or the requests for plain films in the x-ray department, or the number of requests for an endoscopy or the number of letters being dictated and waiting to be transcribed by a secretary.

Demand = capacity, no variation

In the first model (demand analysis.xls), let us assume that we have a constant demand of 10 patients or requests for the clinic or service each week (demand max = 10, demand min = 10). The demand does not vary and is a constant 10. The maroon columns in the model show a constant demand of 10 patients.

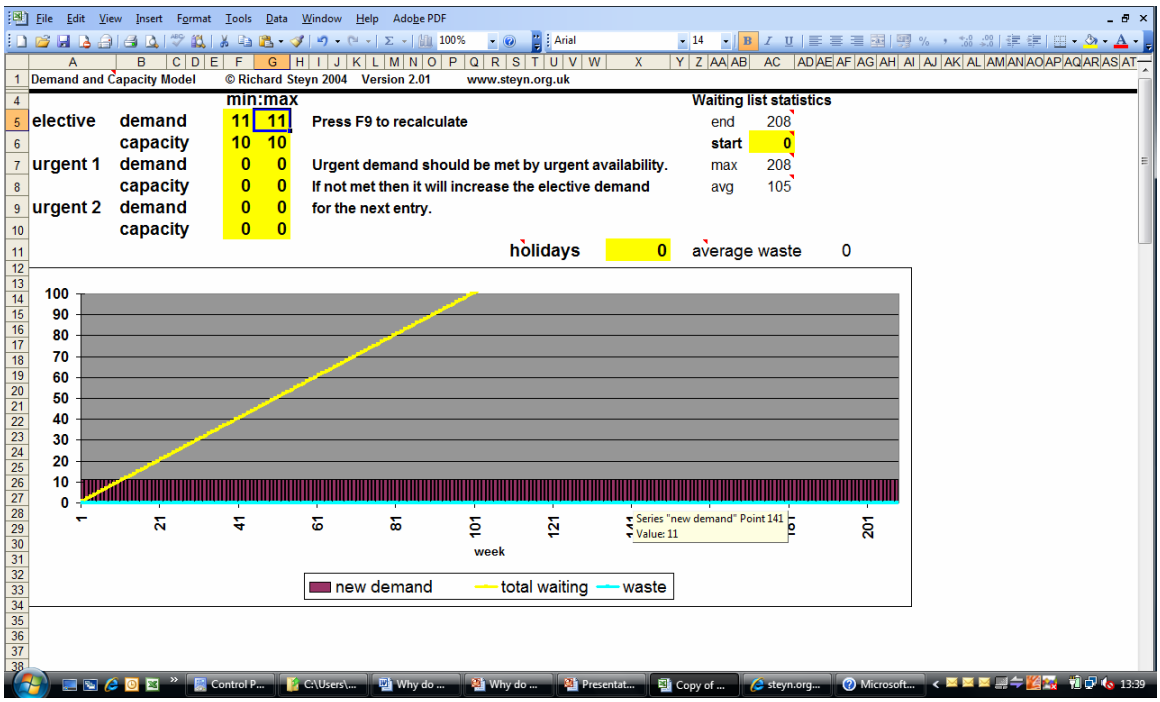


In this case the capacity has been set at 10 clinic appointments each week. (Capacity max 10, capacity min 10)

As capacity matches demand perfectly we will have no patients waiting (no yellow line) and no wasted appointments in the clinic (no blue line – see later).

Average Demand exceeds Average Capacity

If the demand exceeds the capacity by just one patient a week (demand max 11, demand min 11), e.g. we have demand of 11 and a capacity of 10, the queue or waiting list will grow by 1 patient a week (yellow line).



Very few queues or waiting lists grow continuously like this. However the queue for follow-up appointments in some organisations is growing like this since patients are not discharged from care but are requested to come back again and again. Some NHS organisations are finding that it is not possible to book patients in for their follow-up appointments in time. Where they have serious pathology that does need regular follow-up, this a serious and risky issue.

In this case it is vital that the providers of the service record the number of requests (demand) for follow-up appointments and the number of patients currently in follow-up and plan their service appropriately. Can patients be discharged back to their GP? Can the service be expanded to meet the growing demand e.g. for routine repeat endoscopies? Can the patients do their own follow-up e.g. monitoring hypertension or blood glucose at home?

Monitoring the waiting list numbers.

It is very important to monitor the total number of patients waiting at each step in the service every week, and not just the waiting time. If the waiting list numbers are growing it suggests that average demand exceeds average capacity.

However there is another reason why there can be a persistent queue or waiting list.

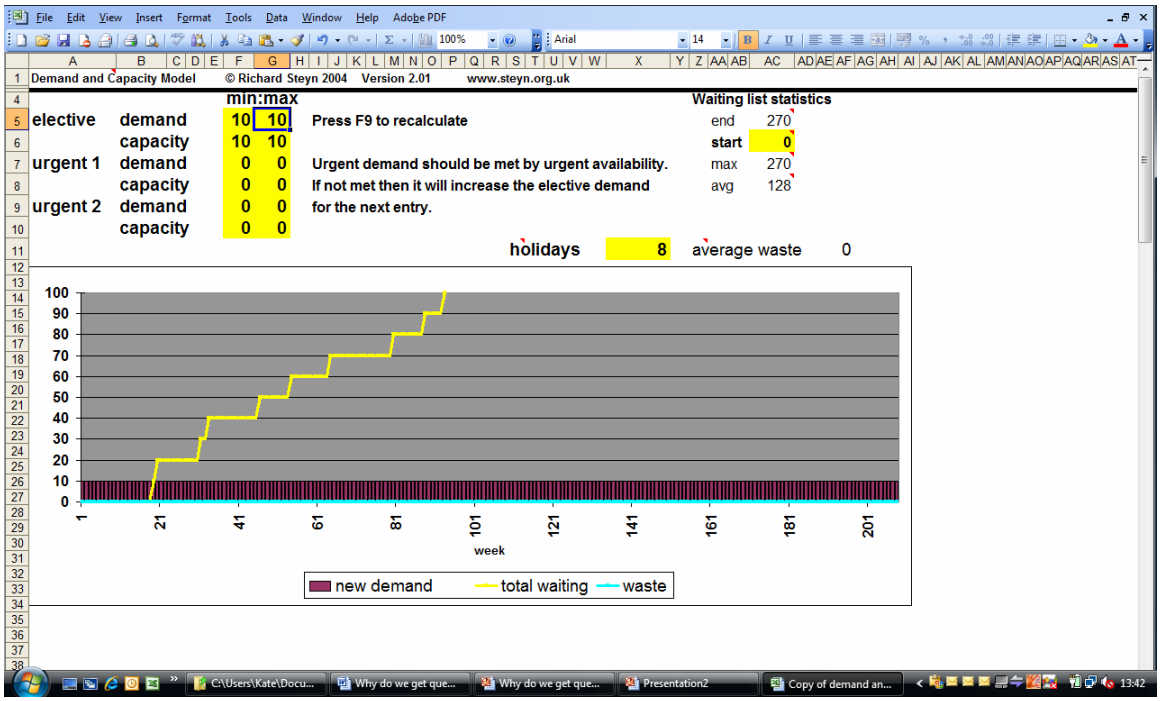
Impact of variation on queues and waiting lists.

Holidays.

We all know that the capacity for a service will vary as a result of holidays.

So in the next example, let us consider that the demand is a constant 10 each week (demand max 10, min10), and the service is planning to provide a constant 10 appointments (capacity max 10, min 10) each week. However if there are to be 8 weeks holidays a year then this will result in the waiting list growing in 'leaps' every time there is a holiday.

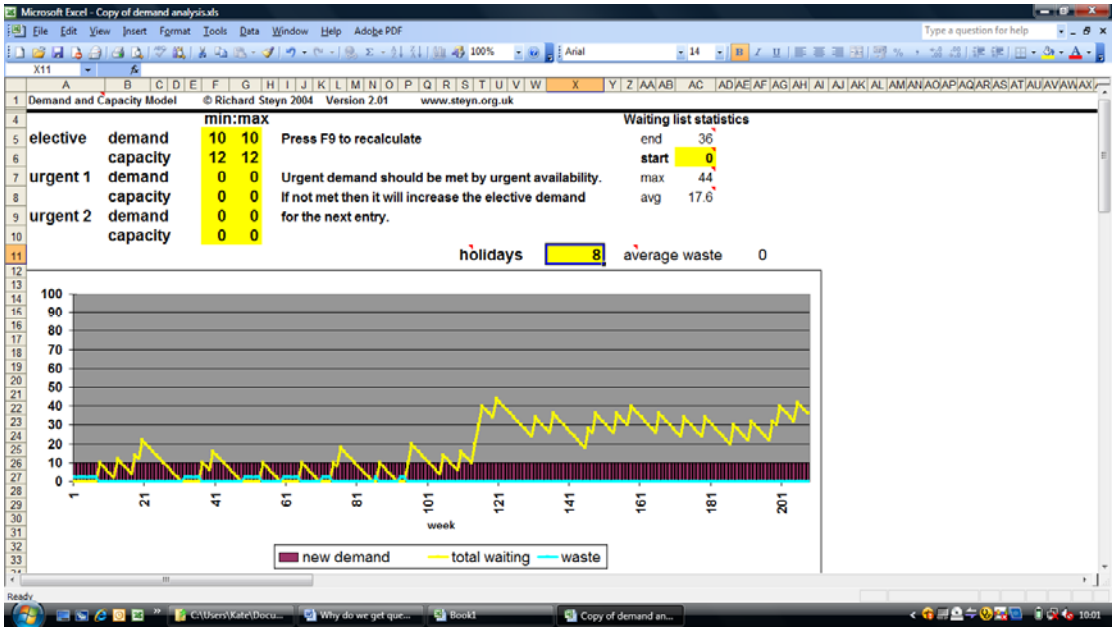
Holiday1



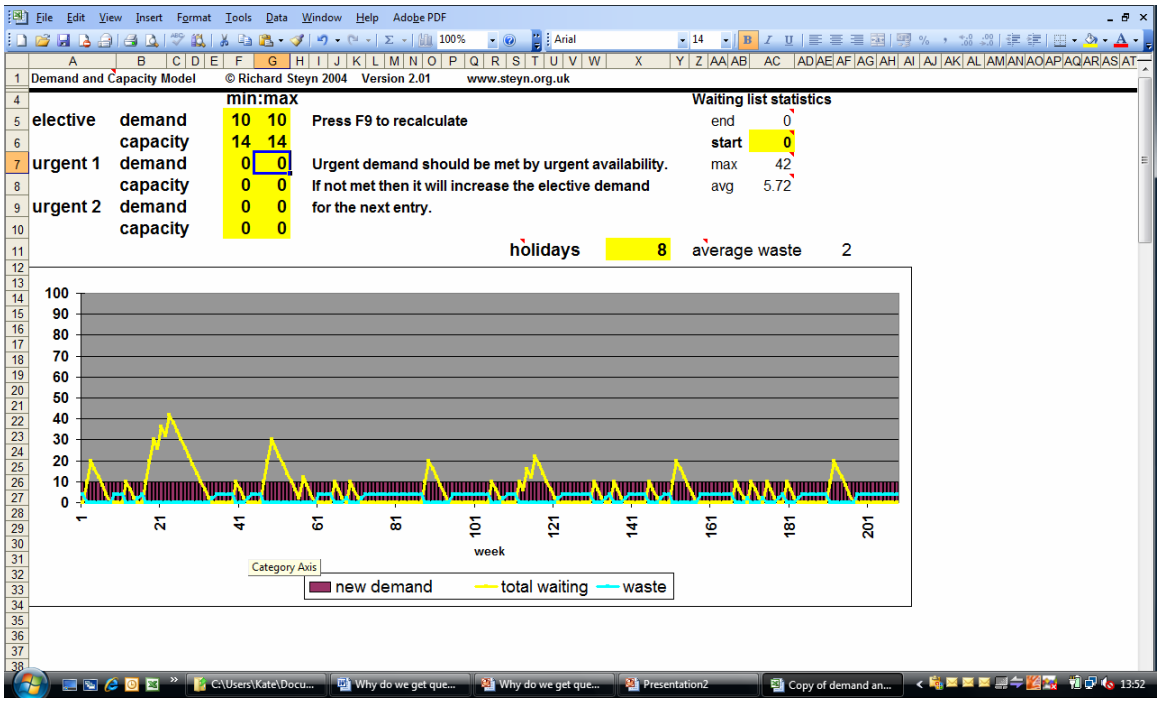
Most services try to compensate for the holidays by increasing the number of appointments available. For example NHS managers, who have no view of their demand, often use the expected or past average activity to plan the service. For example if the expected output of the service (activity) is to be 10 x 52 = 520 patients a year, and there are to be 8 weeks annual leave, then there will be 44 effective weeks rather than 52 (52 weeks – 8 weeks holiday). So the number of appointments they aim provide is 520/44 = 12 appointments each week.

Holiday 2

In this example we will provide 12 slots each week (capacity max 12, min 12) to make up for the weeks when the service is on holiday. In which case the queue will appear during the holiday and gradually disappear after a holiday break. Notice how there is also are few 'wasted' appointments appearing too (blue line). Also notice how, in this example, despite having compensated for the holiday, after a while the queue persists.



To prevent the queue from persisting we have to add in more capacity than the calculation based on averages would suggest.



Putting in a capacity for 14 appointments each week allows the server to get rid of their backlog when they get back from holiday but there are more wasted slots. However the queue is kept under control.

As we have just demonstrated, planning based on averages like this is fundamentally flawed. The rest of the training guide explains why.

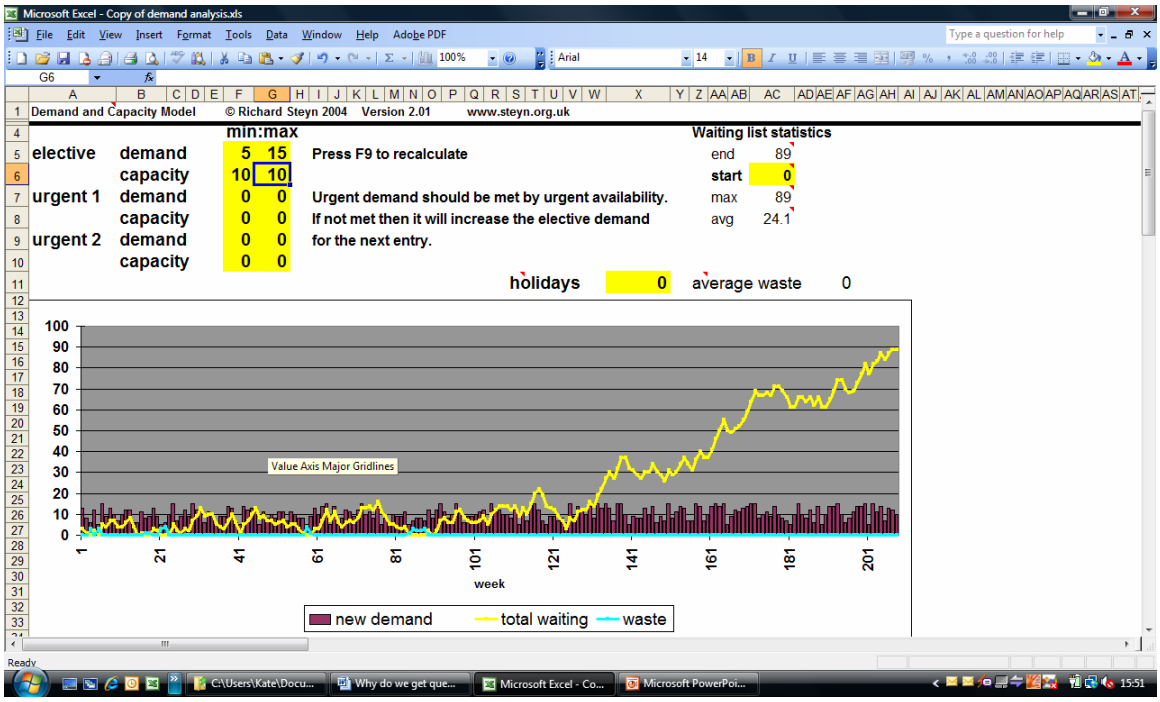
Not only do we have variation in capacity caused by holidays and sickness etc, the demand for a service also varies.

Variations in demand.

In this example we will keep the capacity constant but the demand will vary as in real life he number of patients presenting or being referred to the service is not constant. The same number of patients do not fall ill each week.

Let us consider what happens if the average demand (10) = average capacity (10) but the demand is varying randomly, from 5 to 15 patients a week (demand min = 5, max = 15).

Model: Average Demand = Average Capacity but demand varies



Despite the average demand = to the average capacity a queue or waiting list develops.

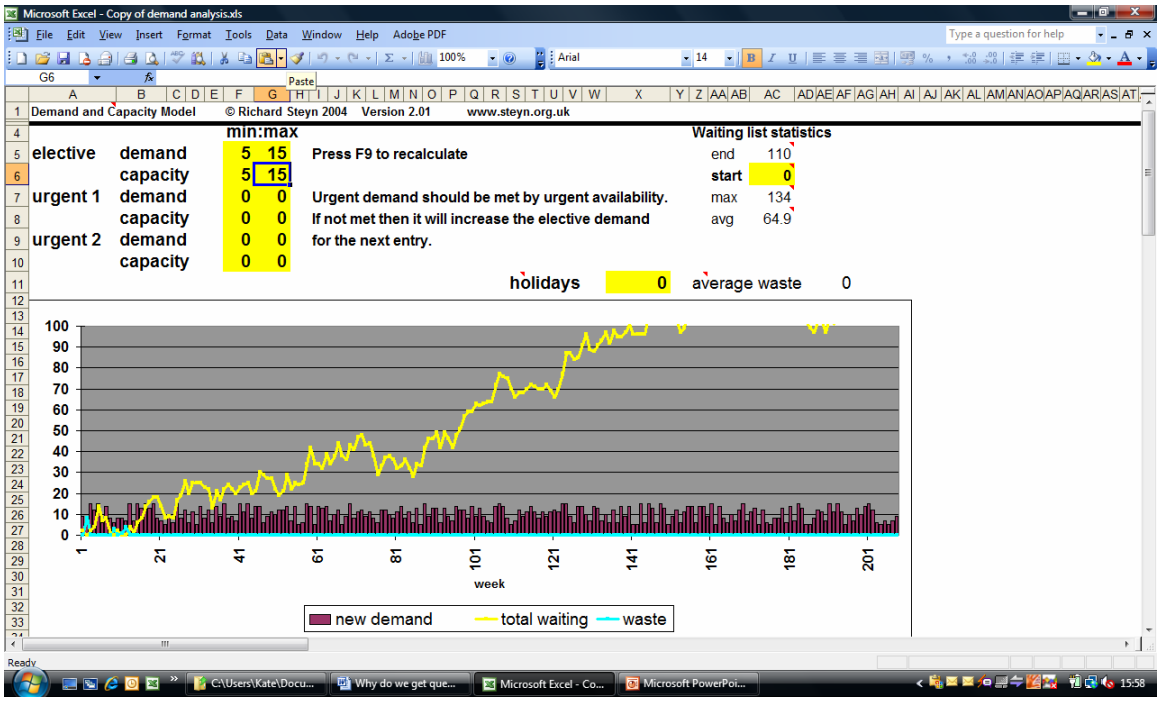
Keep resetting the Excel model by pressing F9 several times and you will see that there is always a queue. However sometimes the queue can be 'okay' but sometimes it can be very long and persistent. These different results are entirely due to the random numbers being generated by the computer System.

Pressing F9 and resetting the model back to 0 is the equivalent of doing a waiting list initiative. Waiting list initiatives are a perennial feature of the NHS. These are periods when staff are asked to provide additional short term capacity to reset the waiting list back to the current target waiting time.
As we can see, despite repeated waiting list initiatives (pressing F9) the queue returns beyond whatever target has been set.

Variations in Capacity

If the capacity in the clinic varies too because of holidays, meetings, sickness etc, (capacity min = 5, capacity max = 15) but the average capacity is still 10 and still equal to the average demand, then the queue or waiting list will be more 'unstable'.

Model: Average Demand = Average Capacity but demand and capacity vary



Keep resetting the Excel model by pressing F9 several times and you will see that sometimes the queue can be 'okay' but sometimes it can be very long and persistent. These different results are entirely due to the random numbers being generated by the computer.

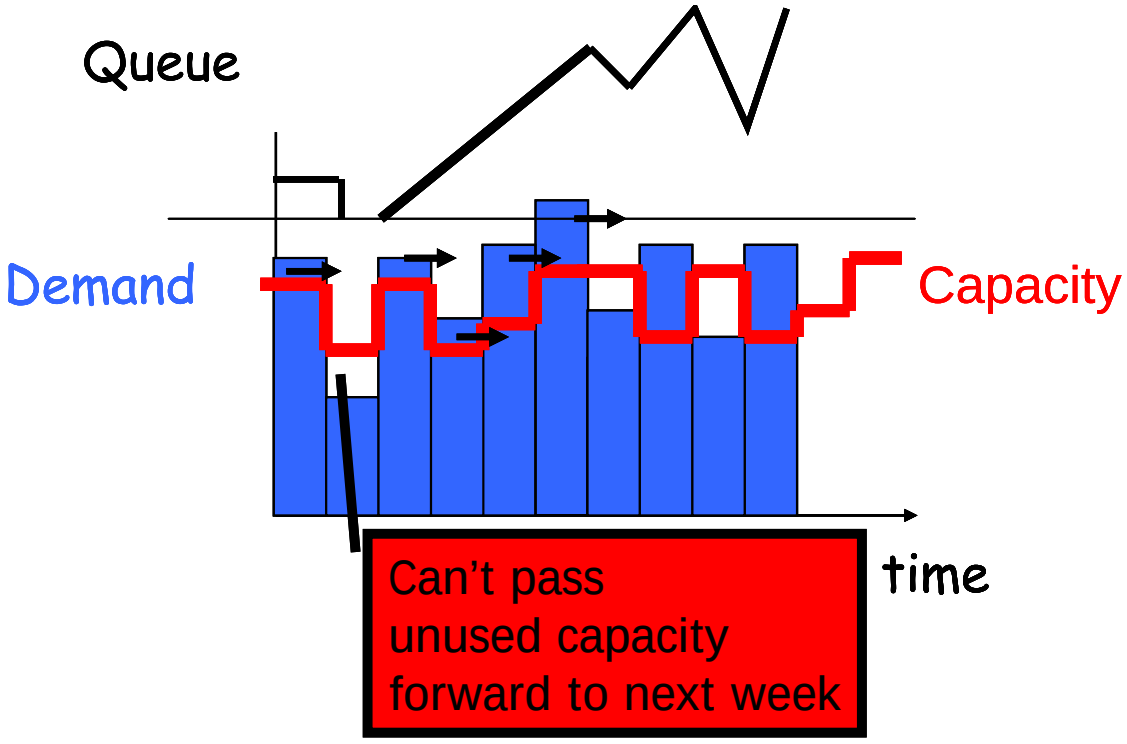
Why is the queue happening?

If the demand is varying and the capacity is varying, then every time the demand is greater than that week's capacity, the extra demand will be carried forward to next week as a backlog or waiting list or queue.

If the capacity is greater than the demand that week (more appointment slots than requests) then the unused appointment slots cannot be carried forward and are lost.

So if we use the average demand and the average capacity we will always get a queue due to the lost capacity slots.

Diagram to show demand and capacity variation mismatch



Keeping Utilisation High.

Once there is a queue for a service, i.e. the patients are present and waiting patients or work can be pulled in 'at short notice' to fill any unused slots. However this is more difficult, inconvenient and often not possible if the patients are in a waiting list at home or at work.

Many western managers believe that having a queue is 'efficient' as it allows high utilisation of resources. However they do not take into account all the cost of waiting e.g. waiting rooms, management costs of the queue or waiting lists, delays to timely care etc.

The crucial difference between Manufacturing and Services.

Traditional manufacturing management uses the periods when the demand is less than capacity to build product and store it in a warehouse for the times when demand is greater than capacity. However, there won't be exactly the right product in the warehouse to serve the future demand, but as long as the products are all the same e.g. a black T model Ford, or 'vanilla flavoured', a good salesman will convince the customer that the product that is available will do.

However in services the product is a unit of time called an appointment to which resources have been attached. Appointments have no shelf life. It is not possible to make and store spare appointments in the warehouse to serve a future queue. So using the traditional manufacturing paradigm that utilisation of capacity by building product for future demand as *the* measure of efficiency is fundamentally flawed.

Tachii Ohno, the founder of the Toyota Production System realised this too. Customers wanted more than a black T Model Ford. If he followed the traditional manufacturing rules of efficiency he would be making varieties of products that customers didn't want. This was pure waste. So he changed the manufacturing management rules and treated his business as if it was a service - manufacturing a car Jus- in-Time to meet the demand.

However, like many western companies, the NHS and other healthcare organisations have not realised this crucial difference. They treat the queue of patients as a means of maintaining a high utilisation of resource. They then spend a fortune on managing the queue rather than seeing and treating patients straight away.

Managing the queue.

Once there is a queue, managers and clinicians have a number of choices:

- Request more capacity
  - Short term
  - Long term
- Prioritising care
- Attempt to reduce the demand

Request for more capacity.

Short term increase:

Waiting list initiatives are requests for short term increases in capacity. These are initiatives when clinicians and other staff are paid extra to work overtime and at weekends to clear the backlog.

We can mimic a waiting list initiative by pressing F9 on the model and this resets the queue to zero, where zero can be the desired or target waiting time. Sometimes the waiting list initiative appears to have worked and sometimes the queue comes back with a vengeance. In which case managers and clinicians often blame 'the demand' i.e. patients or their general practitioners (GPs) for increasing demand in response to a shorter waiting time. However the demand has not changed.

All the evidence so far form those services that are designed to meet healthcare demand (Murray et al 2002) is that demand actually goes down if the overall waiting times come down. If there is a short and consistent waiting time patients and their GPs are more confident that they can access a service when they need to and don't book appointments 'just-in-case'.

Long term increase:

Business cases are written 'for more resources' because of the persistent backlogs. Business cases for more capacity take a long time for approval and implementation.

The majority of business cases fail to understand the reason for the persistent backlog and assume that the backlog is due to the average demand is greater than average capacity. In addition business cases are based on past average activity and future 'estimated' average activity. Activity is a crude measure of the effective capacity not the demand. Very few business cases are based on actual demand and variations in demand. So the majority of business cases that call for more capacity are fundamentally flawed.

Prioritising & Carve Out

In the face of a persistent queue or waiting list, (and while they wait for their business case to be approved) managers and clinicians have no choice than to prioritise patient on the basis of 'their clinical need'.

Prioitisation has grave consequences:

- it leads to extra 'prioritisation' steps and delays in the process
- it wastes resources that could otherwise be seeing patients.
- and leads to very complex booking rules that very few, if anyone, understands.
- often it is not possible to establish the clinical urgency on the basis of the referral so the risk in the System is increased

Clinicians and mangers then 'carve out' the available resource into different 'buckets' reserved for different patient types. Sometimes this is called ring-fencing or 'batching'.

Impact of carve out

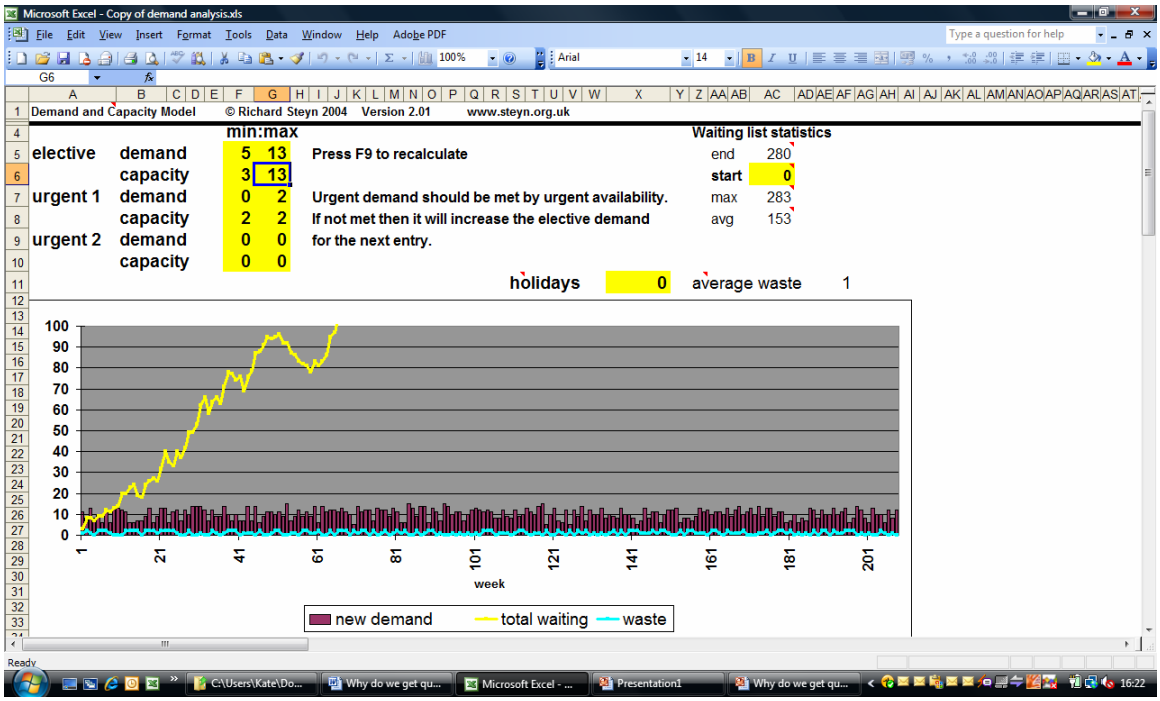
So in the model demandanalysis1.xls we are now going to split the demand into sub-groups of patients with different clinical priorities. This is called carve-out.

So for example let us say we can have up to 2 patients with suspected cancer referred each week but we don't always get a suspected cancer patient referred to us. In this case we will show the demand as Urgent 1: demand min =0, max =2. So we will reserve 2 slots for the cancer patients Urgent 1: capacity min 2, max 2.

The overall demand is still varying from 5 to 15 around an average of 10.

So we need to ensure that we haven't changed the overall demand and capacity. So the main demand min: 0 max 3 and the main capacity will be will be 3 and max 13.

Model showing 2 queues for the same service:



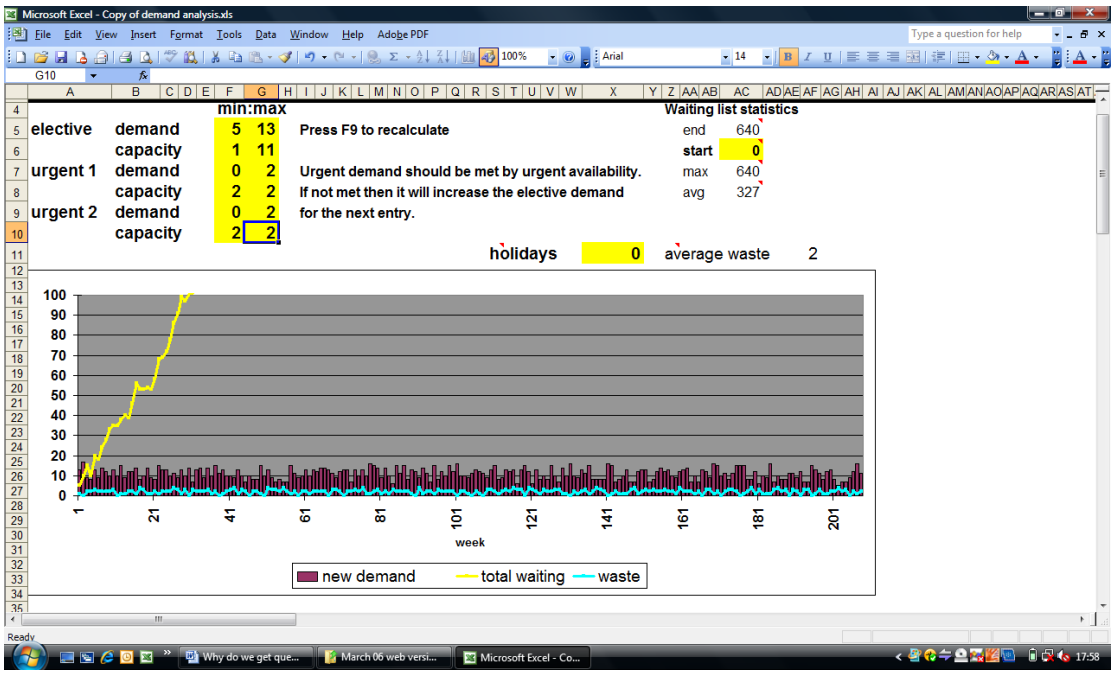
In case you are having difficulty in understanding how to work the demand and capacity for multiple queues, here is an alternative way of looking at the yellow squares in Richard's model:

	Demand min	Demand max	Capacity min	Capacity Max
Elective	5	13	3	13
Urgent 1	0	2	2	2
Totals	**5**	**15**	**5**	**15**

What happens now is that the queue gets worse and we have wasted slots. This is because the cancer appointments will be wasted if we don't have a suspected cancer referral that week.

We can then carve out more slots, for example, children or other vulnerable groups. Again the overall demand and capacity hasn't changed, but we are now wasting more slots due to the carve out.

Model showing 3 queues for the same service:



	Demand min	Demand max	Capacity min	Capacity Max
Elective	5	11	1	11
Urgent 1	0	2	2	2
Urgent 2	0	2	2	2
Totals	**5**	**15**	**5**	**15**

In such circumstances clerks and secretaries are often juggling the queue and asking patients to come in at short notice. Patients may receive multiple cancellations and re-appointments. Clinicians notice the wrong type of patient in the clinic slot and make the booking rules even more difficult to follow.

Tempers become frayed; GP, administration staff and even patients are blamed for the errors and chaos reigns.

Examples of carve out.

Orthopaedic clinic

This clinic timetable shows the different types of carve out;

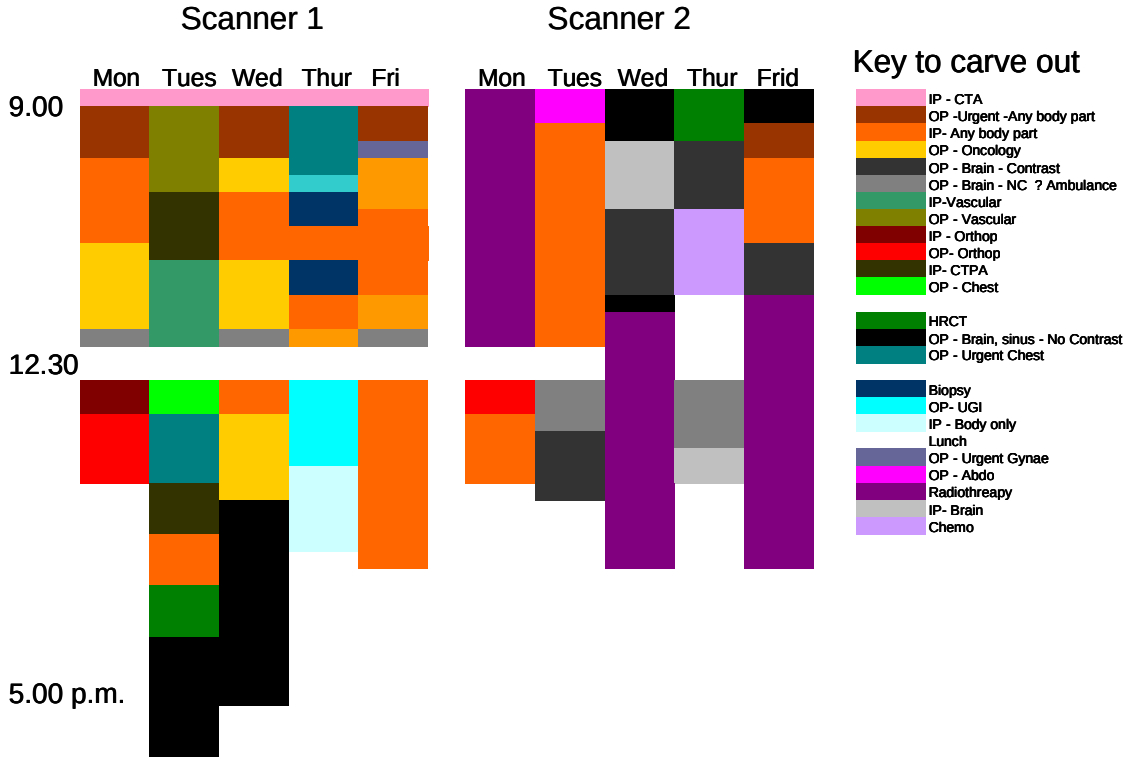
fracture clinic: follow up fractures,
ortho new orthopaedic referrals from GPs
daily fracture clinic for new fractures

		MONDAY			TUESDAY			WEDNESDAY			THURSDAY			FRIDAY		
		FRACTURE	ORTHO	Daily #	FRACTURE	ORTHO	Daily #	FRACTURE	ORTHO	Daily #	FRACTURE	ORTHO	Daily #	FRACTURE	ORTHO	Daily #
AM	CONSULTANT	0.5	0.5		0.5	0.5		1			1		1	0.5	0.5	1
			1	1	0.5	0.5	1	1		1	1					
	REGISTRAR	0.5	0.5		0.5	0.5		1			1			0.5		
			1		0.5	0.5		1								
PM	CONSULTANT		1			1		1	0.5		trainign and audit				1	
			1			1										
			1													
	REGISTRAR		1			0.5										
			1			1										
Sessions		10			8.5			6.5			4			3.5		

The result is that the capacity as measured in sessions (a session = 3.5 hours) varies thought the week. Whenever there is a bank holiday Monday, 30% of the clinic capacity for that week is lost. In addition to this carve out though not visible, the consultants have different sub-speciality interests too: knees, shoulders, hands etc. On top of this there are urgent , soon and routine appointments.

CT scanner

This timetable shows all the different kinds of appointment slots that have been carved out for 2 CT Scanners in a radiology department:



It is impossible to ‘balance’ this number of queues. Is it any wonder that there are persistent queues for these services?

Summary so far:

We have seen that there are 3 basic reasons for a queue or waiting list

1. average demand > average capacity
2. average demand = average capacity, but demand and capacity vary and there is a mismatch between the variations
3. having a queue to keep the utilisation of a resource artificially high.

Since services are planned on the basis of averages, the majority of queues in the UK NHS are due to the mismatch between the demand and capacity variations.

In response to the queue or backlog managers and clinicians call for more resources and prioritise or ‘carve out’ the available capacity to reserve slots for their more vulnerable groups of patients.

This make the waiting time worse, increases staff & patient stress and increases risk and cost.

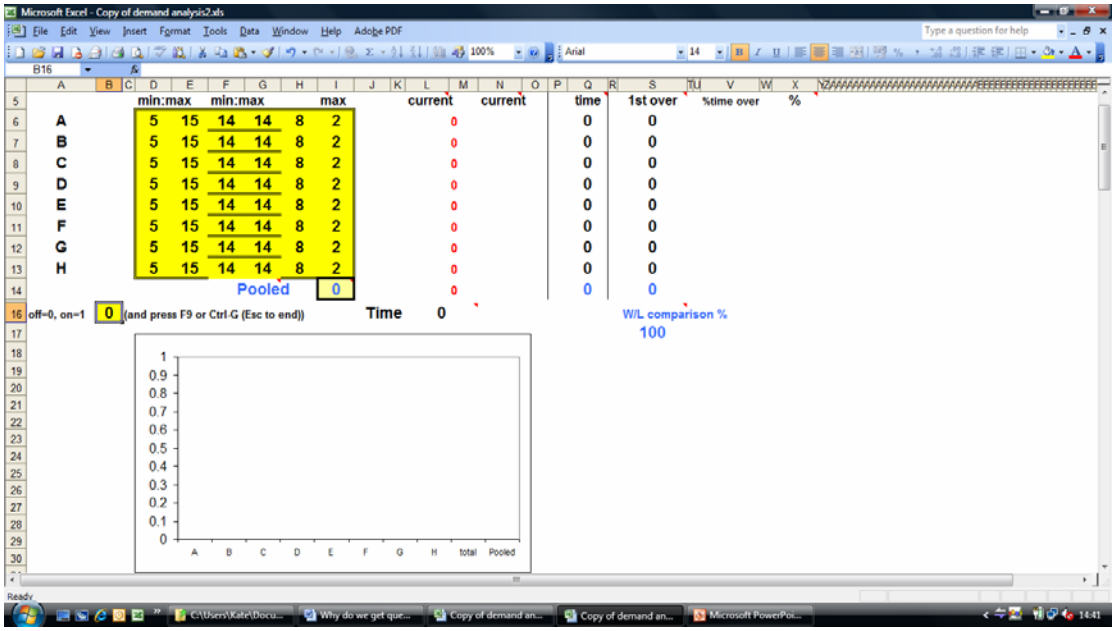
So what should do we do instead?

The first thing to do is to acknowledge the problem and to start reducing the variations in capacity. This is in our control.

We can reduce the Carve Out and pool resources so that the impact of holidays and sickness is minimised.

Pooling

Please use the model demandanalysis2.xls to show the impact of pooling the capacity.



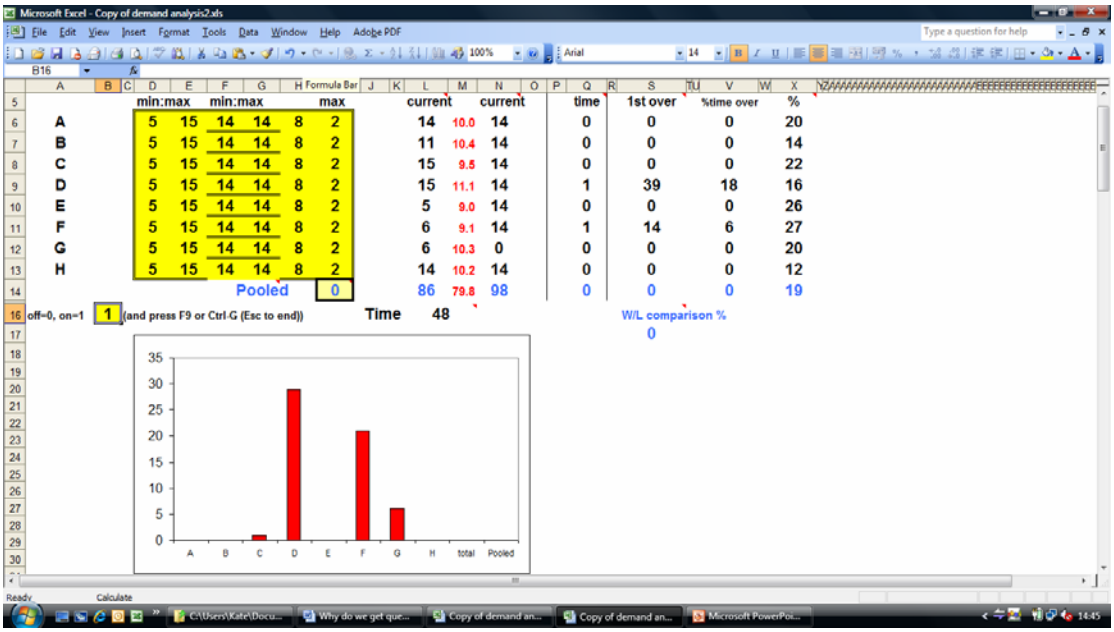
In this model we have got 8 ‘servers’ providing a service (Servers A to H). This could be 8 GPs, 8 secretaries, 8 ambulances, 8 chiropodists, 8 midwives, 8 surgeons, 8 laboratory technicians, 8 social workers etc.

Whatever the service is are, each server receives a demand of 5 to 15 patients, averaging 10 per week. Each has got 8 weeks holiday a year which they take at random with no reference to when their colleagues are away. To adjust for the holidays we have put in 14 appointment slots in each week.

In order to avoid the carve out problem, we want all patients, whatever their clinical condition, to be seen in 2 weeks. This means that we can only want a maximum of 14 x 2 = 28 patients waiting.

To turn the model on, put 1 in the yellow box marked off = 0, on = 1 and then press control G

The number of patients waiting at each server will show as a red bar.



As we can see over time, the queues for each server come and go, despite the fact that the demand and the capacity is the same for each server.

‘Performance management’.

Pretend that it is time for a performance review and take a snapshot of the servers current position by stopping the model. (Press ESC twice).

Performance managers can then make ‘judgement’s about an individual’s performance. In the screen shot above, server B might be judged as a ‘good performer’ as he or she has no queue, but D & F are ‘poor performers’ because they have a long queues.

Clinicians and managers fail to recognise that the demand and capacity for all 8 Systems are the same and that the length of the queues is entirely due to random mismatch of the variations. This doesn’t stop statistically illiterate performance managers making inappropriate judgements about individuals behaviours!

Start the model again (control & G) and wait a bit. The queues will come and go. Pretend it is 6 months later and time for another ‘performance review’.

(Stop the model gain by pressing ESC twice).

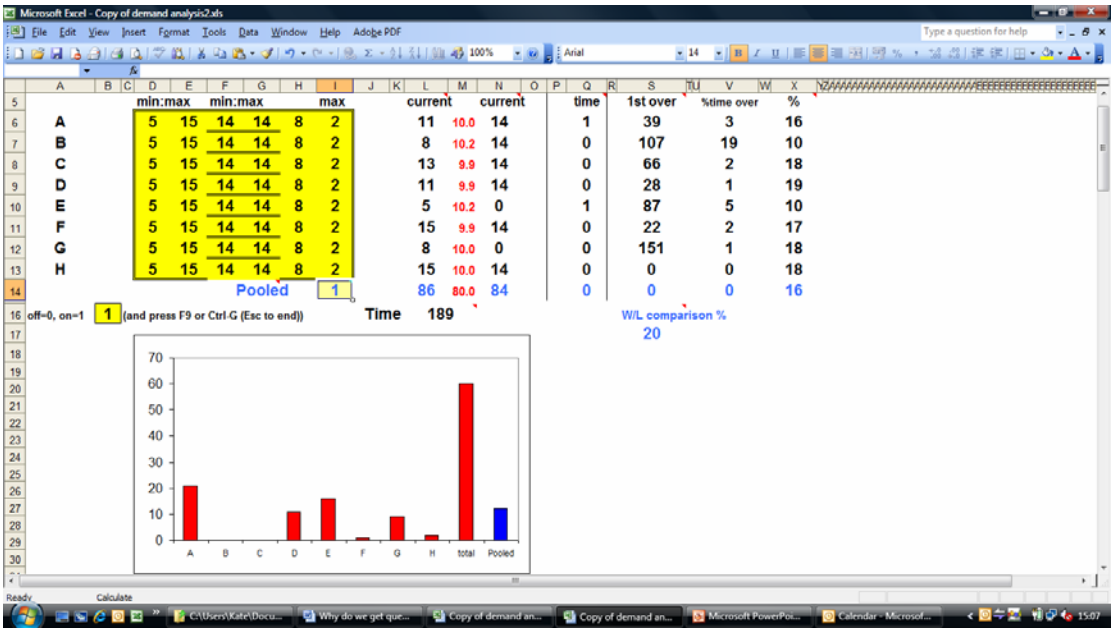
Some queues will have improved – not because of ‘improved behaviour of the individual’ but due to the inherent variation mismatch and random chance!

One way out of this ‘performance management nightmare’ and persistent bullying is for the servers to pool their demand and their resources. Now if one is way or ill, the demand will go to the next available server.

Pooling Model 2

In the model now turn on the pooled queue by putting 1 in the yellow box marked ‘Pooled’.

The pooled queue has always been running, but now a further 2 columns will appear in the graph.



The penultimate red ‘Total’ column is the sum of all the individual queues if the servers still worked independently. The blue Pooled column is what would happen if all the demand was put into a single queue and took the next available server.

The impact of pooling can be dramatic.

As we can see the pooled (blue) queue hasn’t exceeded 28 patients or 2 weeks wait at any point. Now patients could be booked in when they needed to be with no need for any extra ‘prioritising’ administrative steps releasing wasted clinical resources.

There are many examples of pooling the demand and capacity for services. Many airlines now ask customers to form a single line or queue at check in desks. The queue looks daunting but is processed very quickly. In contrast super markets are constrained by the size of shopping trolleys and still have multiple queues, one for each counter.

Specialisation.

The argument against pooling a service is that it is no longer possible for clinicians to be able to deal with every clinical condition that is sent to them. So what do we do about specialisation?

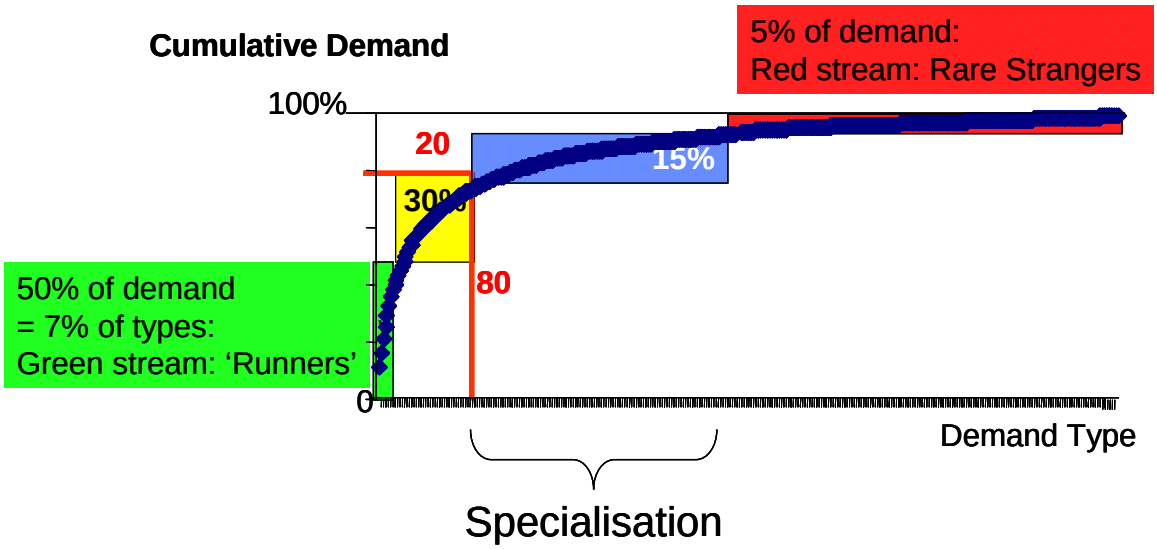
We are not advocating that the whole service should be pooled so that gynaecology patients are seen by ophthalmologists, but that within a specialist service it is possible for many of the patients to be pooled.

Pareto Principle.

If we look at the different types of patients being referred on the basis of their presenting complaint or clinical condition and the cumulative volumes of patients presenting with these conditions, then we discover that the Pareto Principle applies. Vilfredo Pareto (1848 to 1923) was the Italian economist who noticed that the 80:20 principle: 80% of country's wealth is owned by 20% of individuals. The Pareto analysis is a useful tool and the Pareto 80:20 principle is found to apply to many things, especially the demand for healthcare.

A Pareto analysis helps define should and shouldn't be pooled. The Pareto can be broken down into 4 main sub groups that require different operational management. (Ian Glenday)

Pareto Analysis:



50% of patients will have 7% of the clinical conditions (green stream) and all servers should be able to deal with these. The process for these should be standardised on the basis of the best evidence based practice. Instructions and training should be visible so that all staff, and indeed the patients with these conditions, should know what to do.

30% patients will have the next 15% of conditions and any specialist should be able to diagnose and treat these conditions (yellow stream). The next 15% of patients will have rarer conditions (blue stream), that all specialists should be able to start the initial investigations passing them on to who ever specialises in the treatment of each of these conditions once the diagnostic tests have been requested.

The issue comes with the 5% of patients (red stream) who have the very rare conditions. Since there are small by highly variable volumes and processing times for these patients, they should be separated out so that they do not disrupt the predictable green stream. In many cases once they have been diagnosed, these patients are referred elsewhere for treatment. The problem for teaching hospitals is that they are asked to deal with a normal general hospital's demand and added into this is the rare 'red stream'. This is why each department should do their own Pareto analysis to work out how their team should best deal with their demand.

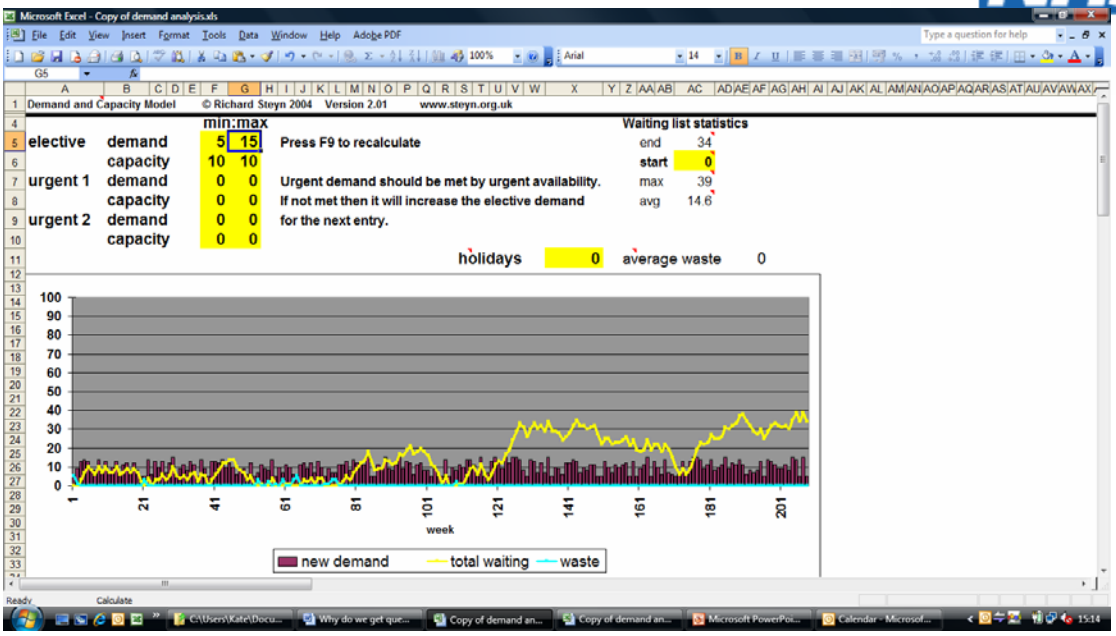
Sorting their Pareto is an important task for the clinicians to do, as it allows for appropriate pooling and sub-specialisation. It also gives them an opportunity to standardise the work for the 50% of patients with the common7% of common conditions. This allows junior and less experienced staff to help with 50% of the work.

Having identified what the common things are that occur commonly, this 50% of demand for 7% of conditions varies the least and these appointments and resources can be scheduled evenly every day or week making the service much easier to run. This is Level Scheduling and, often, the economies of standardisation and repetition allow 50% of the work to go through quickly, often requiring only 30% of the total capacity.

So where have we got to?

We have removed the carve out and pooled appropriately. So we have smoothed the capacity (capacity min =10, max =10) and the System is much more stable, but we are still left with the problem that our average demand is equal to our average capacity and we have a persistent queue.

Return to Model demandanalysis.xls



What do we do next?

Reducing demand: this does not work.

The response of most managers and clinicians at this point is to attempt to reduce average demand by ‘demand management’. Complicated referral guidelines are developed and often complicated ‘triage’ services’ are set up in primary care to filter the demand ‘appropriately.

This is flawed approach for many reasons:

1. Is it possible to ‘prioritise’ or filter the demand more appropriately in primary care without diagnostic equipment?
2. How many people are required to filter the demand?
3. How many filtering steps are involved and what impact do these additional steps have on the time to treatment, the demand and the variation in demand?
4. How do the ‘demand managers’ know they have got it right or wrong? What feedback is in the System? What redress do patients or their GPs have if a case was rejected inappropriately?
5. Patients will find their way into the System somehow to get sorted out.

The important role for ‘demand management’ is not to try and reduce overall demand but to direct the patients into the appropriate service.

The service needs to monitor the demand coming through and feedback based on the Pareto principle: concentrating on managing that 50% of patients who are referred with the 7% of conditions.

Increasing Capacity.

If we have an unpredictable variation in demand, and we don't want a queue, we have no option but to increase capacity. But by how much?

In order to plan capacity appropriately, it is vital that we understand:

- the average demand,
- the variation in demand and the pattern of variation in demand.

Healthcare organisations will have to learn to measure and monitor demand continuously.

If there is a predictable variation in the demand, e.g. for minor injury services in the summer, it may be possible to vary the capacity to meet it. However it is very difficult to turn 'off and 'on' the capacity of skilled staff.

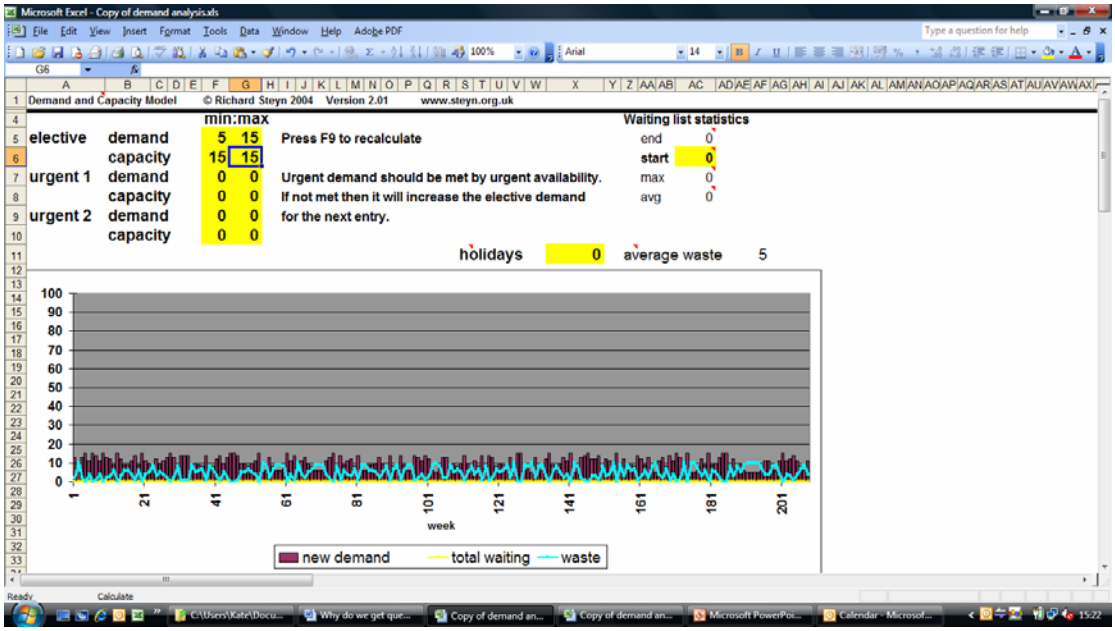
However despite a variable but predictable demand it may be very difficult for the staff to flex the capacity to meet the daily demand exactly. This is the case in Richard's model demand analysis.xls. Though the demand is varying between 5 and 15 randomly and it is not possible to predict exactly what the next demand will be.

In this case, the capacity required will depend on the maximum waiting time that patients will tolerate or can risk.

Emergency response

If we have to see patients straight away then we will have to set our capacity to meet the peaks in demand (capacity min 10, max=10). This would be true for an emergency service. .

Model 1: demand analysis.xls



In this case we will have a lot of wasted appointment slots or capacity but no queue. This is what we would want to see in the resuscitation area of the Accident and Emergency department.

Elective service

If on the other hand we can afford patients to wait a bit, what is the optimum balance between the queue and the wasted appointment slots?

Modelling queues and waiting lists.

This is a huge topic and the optimum balance between the queue length and the utilisation of the resource is dependent on:

1. The variation in demand and the statistical distribution of the variation of the demand
2. The variation in capacity and the statistical distribution of the variation in capacity.
3. the variation and distribution of the variation in the processing time (cycle time) for the different types of patient i.e. the impact of case mix in the demand

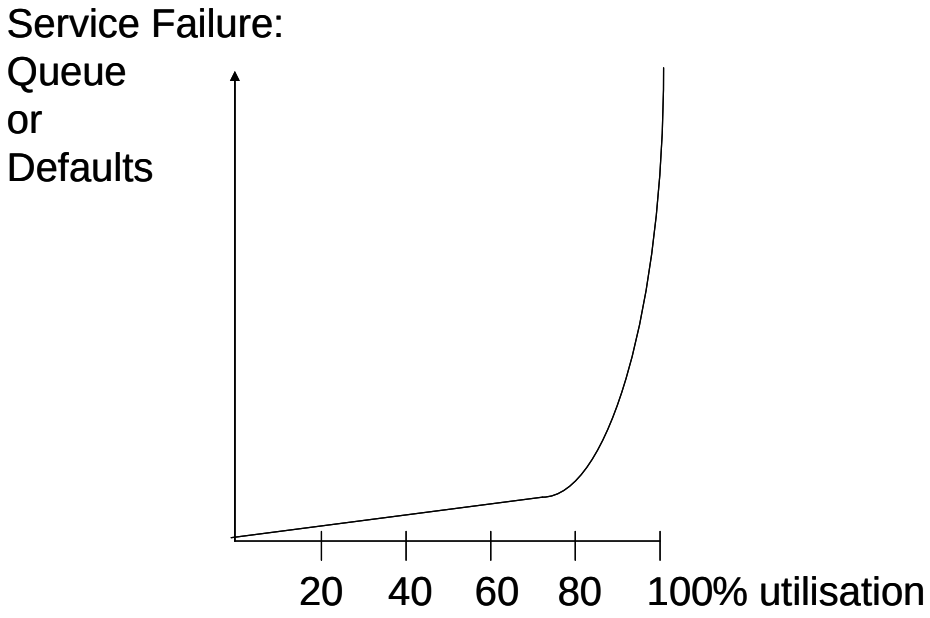
Once these 3 variables have been measured, a Discrete Event Simulation (DES) can be used to give the maximum waiting times at different levels of capacity and utilisation.

In most cases this is not necessary.

Erlang's Rule of Thumb.

Before the advent of computer models and Discrete Event Simulations (DES), manufacturing engineers used a very useful 'Rule of Thumb' to help them address this problem. Agner K Erlang (1878 -1929) was a Danish mathematician who observed the relationship between the utilisation of the plugs on a telephone exchange and the response time. At 20% of the plugs plugged in, the service was very good, with very few customers being asked to wait to be connected. The service remained very good until about 80% to 85% of the plugs were connected, and then the service level fell off dramatically.

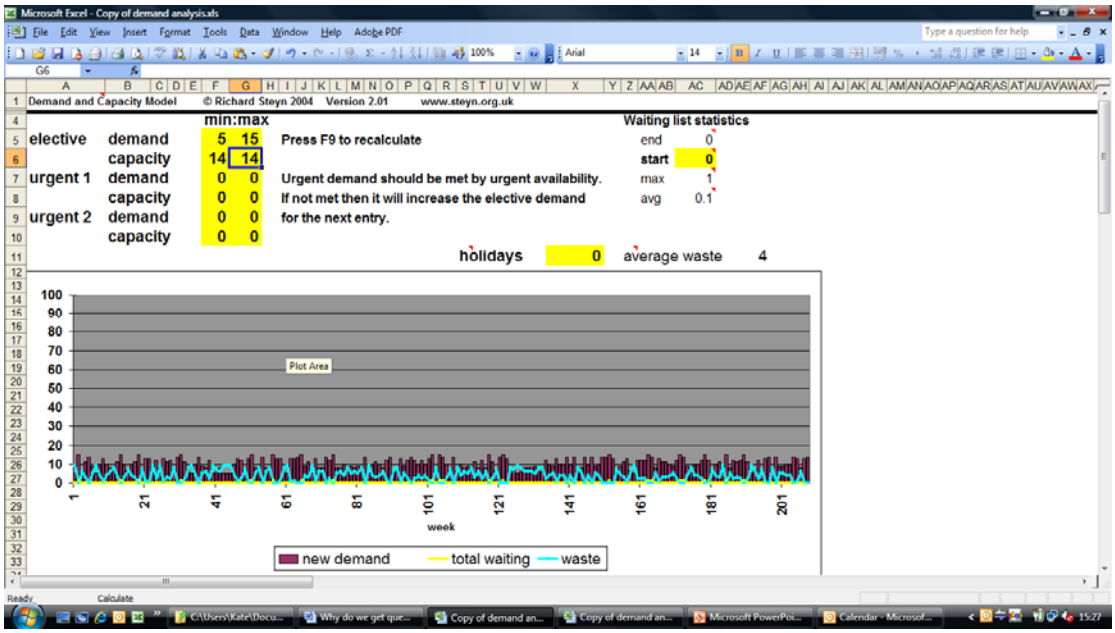
Diagram to show Erlang's relationship between utilisation of a fixed capacity and the service level in a System where the demand is varying



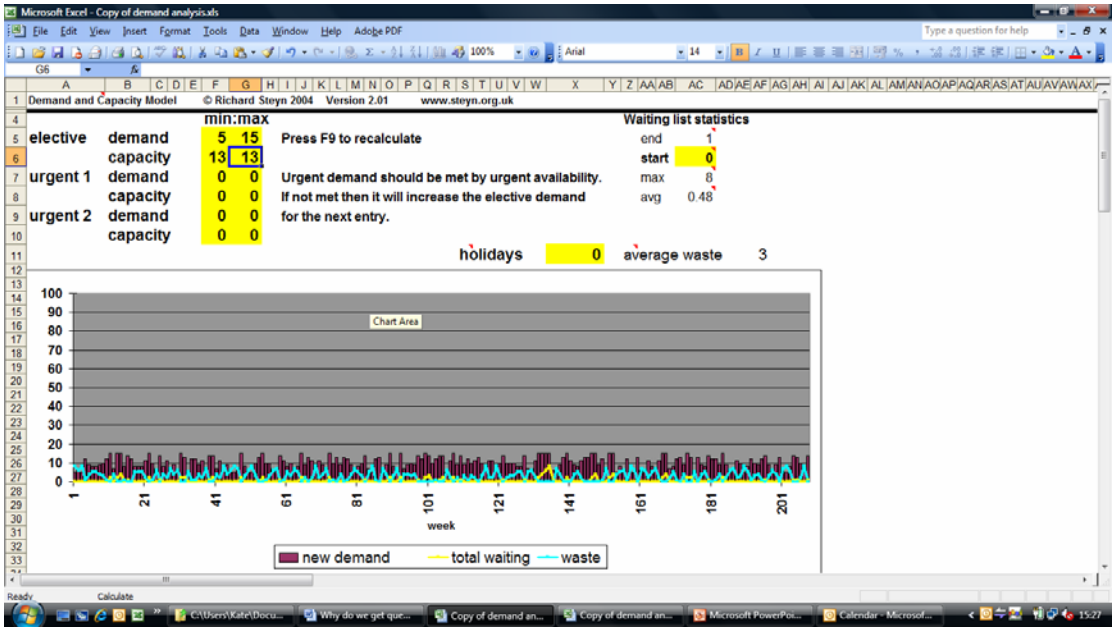
We can illustrate this principle in Richard's Model

Model 1: Erlang's Rule of Thumb

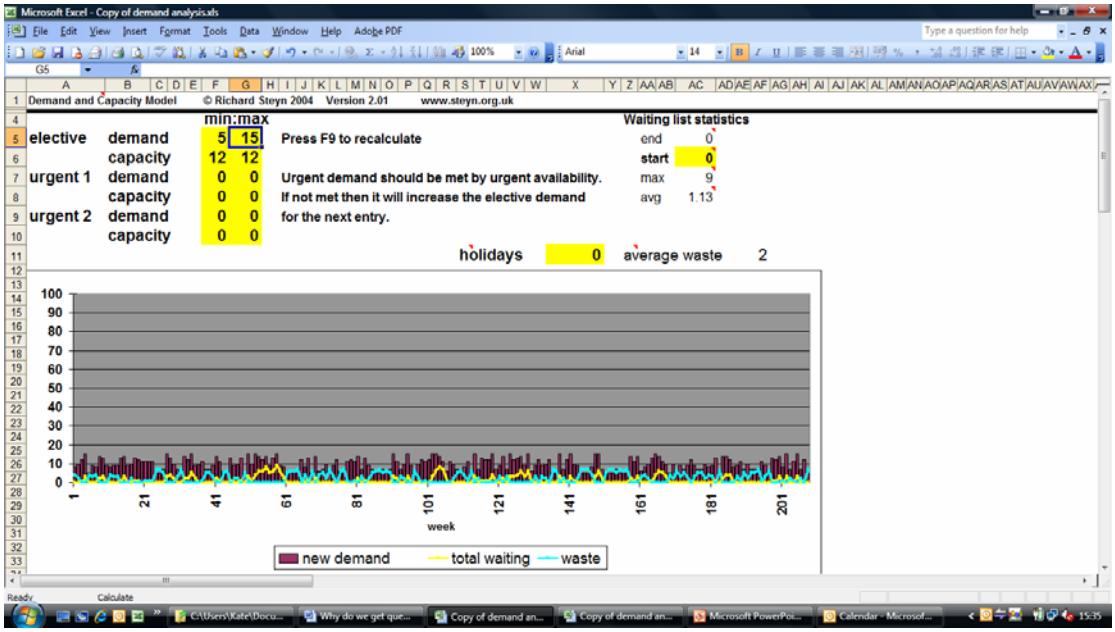
In Richard's model, if we drop the average capacity down from 15, to 14 appointment slots, then most of the time we have wasted slots and just occasionally 1 or 2 patients waiting till the following week. The System is very robust. (Press F9 several times to see how robust the System is)



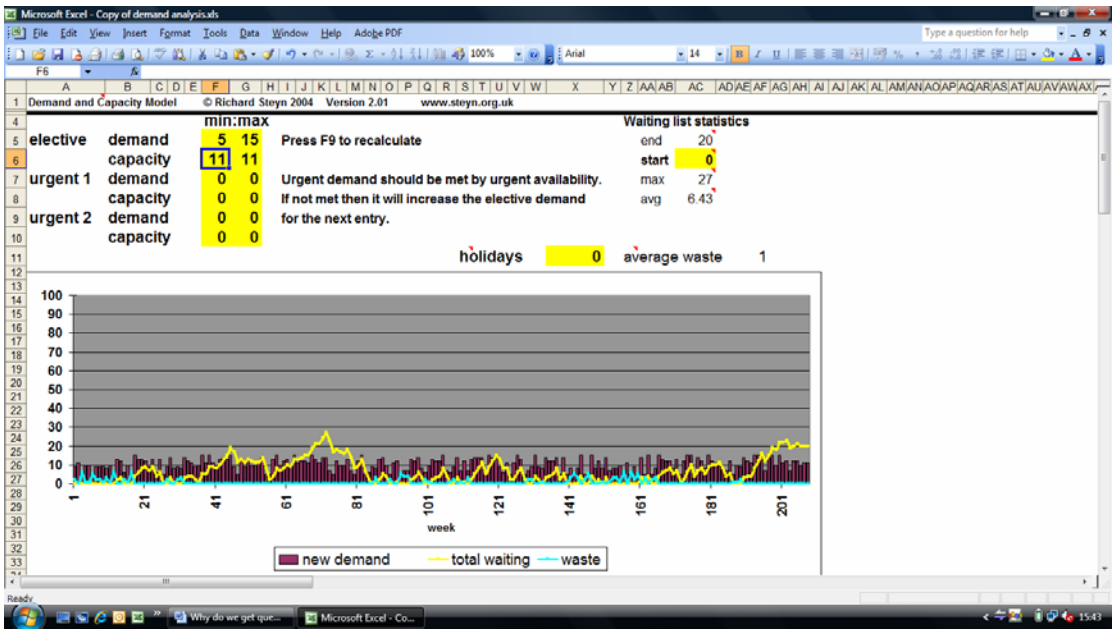
If we drop if to 13 appointment slots, there are less wasted slots, but the queue gets a little more persistent but never out of hand. In this case it is possible for staff to flex their capacity to meet the demand. The server would leave the clinic early on some days but would be quite happy to stay behind occasionally and see an extra patient other days. The System is still very robust.



At 12 slots, the queue is now getting more persistent and we would probably have to prioritise thus setting in place the vicious System of carve out which wastes so much resource.



At 11 slots the System is starting to be very unstable, and at 10 slots, the average, it falls over completely as we have already seen.



Monitoring the demand and capacity for a service

So whether we use a Discrete Event Simulation or whether we go by Erlang's Rule of Thumb (and accept all the error that we will have due to the statistical distributions of the variation in demand and capacity) we need to:

A. Understand the response time required by the patients.
B. Understand the demand:

So we need to measure and monitor each week:

- The demand (number of patients)
- The capacity (number of appointment slots available)
- The activity – the number of appointments actually performed (including the 'Did Not Attend' patients (DNAs) etc)
- The backlog: the number of patients waiting at the end of each week.

Run charts

These need to be plotted as run charts with consecutive weeks on the x axis and the number of patients, appointment slots, activity and backlog on the y axis. It is vital that we monitor the demand and the variation in demand for our service. We can then check with our customers any reasons for changes in demand or capacity of the service.

Normal and Special cause variation (Walter S Shewart)

Variation can be defined in two broad types:

The 'normal or 'common cause' variation is reflected in an 'unchanging pattern' of variation as seen on the run chart. This is due to the random 'chance' events in the process
e.g. the number of patients presenting each week day in general practice

Special cause variation is when the pattern of the 'normal' or 'common cause' pattern of variation changes and / or is assignable to some predictable or special event.
e.g. the number of patients presenting at weekends in primary care or the winter summer cyclical pattern in the number of minor injuries presenting to emergency departments.

By monitoring the demand we can identify and predict the special cause events and match the capacity to meet these. The common cause events require a different strategy.

Estimating Erlang's Rule of Thumb from the Run Chart of the demand

From a run chart (or time series) of the demand we can see where the 80% level of the normal peaks in demand lies and see if we have got enough appointment slots each week to meet this. As a rule of thumb, this is the minimum level of average capacity that an ELECTIVE System can run without the queue growing out of control.

However there is a trap here. Remember that if we need an instantaneous response, then we must set the capacity to meet the peaks. Planning bed capacity is a good example of this trap. If we plan to have only enough beds to ensure that we meet 80% of the variation in demand for beds, we will guarantee a blocked A&E department and cancelled elective operations. The demand for beds requires an instantaneous response.

Reducing Cost

From the example in Richard's model where we had a min demand = 5 and maximum demand = 15 patients, in order to have a stable service, we needed to increase the average capacity from 10 to 13 appointment slots each week. This is a 30% increase in capacity and cost! Most NHS managers would baulk at this, and rightly so. The income will still only be for 10 patients, but this will have to pay for the capacity to provide 13 appointment slots. So now there is a real challenge:

1. will the business be profitable?
2. but if we reduce the capacity we will end up with a waiting list, and all the non value adding costs of managing the waiting list and risks of delay.

So what choice do we have?

Reduce the variation to reduce cost.

The amount of capacity we require depends on the variation in the demand, not on the average demand. What if we could understand the causes of variation in demand and try and reduce it? We would not be turning patients away from care, but understanding why different numbers of patients present or are referred each day. In the majority of cases the demand on one service varies according to the variations in upstream capacity. So if we could understand the variation in upstream capacity we could work to reduce these. This would mean that we would require less capacity to meet the same demand!

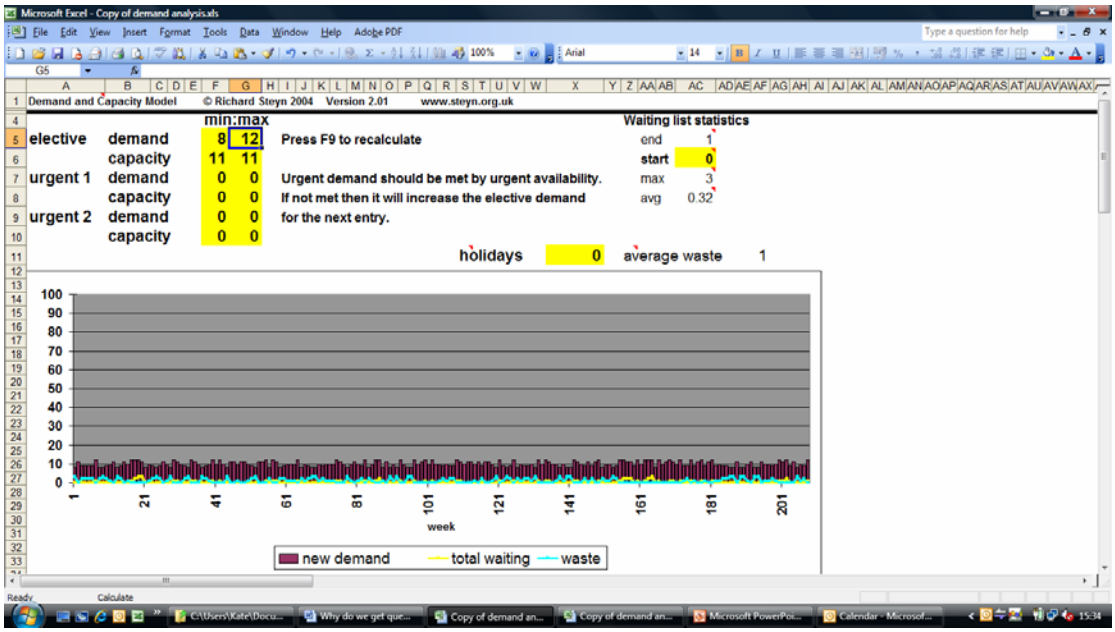
In many cases the variation in demand is very predictable so we could flex our capacity to meet these predictable variations: e.g. minor injuries are greater in the summer.

However some aspects of demand is not always predictable, and we will have to meet this unpredictable variation

Model 1: reducing variation in demand.

So in Richard's model let us pretend that we have persuaded the GPs to reduce the variation their upstream capacity for example buy working weekends. In this case the demand they pass to us would vary less and we would need less capacity to provide the same service.

So If we reduce the variation in demand from 5 to 15 to 8 to 12 (the average demand is still 10 patients) then we only now need 11 appointment slots each week to provide the same level of service.



The waste in the NHS is such that finding an extra 10% of capacity (as int h is example) can be easily be done by applying Lean Thinking principles to redesign processes.

More variation means more capacity is required to keep the queue under control.

The more variation there is in the System, the greater the capacity is required to keep the queue under control. Try experimenting with Richard's model to show this in action.

Demand			
Average	Min	Max	How much average capacity is required to keep the queue under control?
20	18	22	
20	16	24	
20	14	26	
20	12	28	
20	10	30	
20	4	36	

Impact of variation on the total process time.

Remember we are not just dealing with the demand and capacity at one step, but the demand and capacity at every step. So at each step the waiting time will vary for each patient and so the total process time, e.g. Time from GP referral to first definitive treatment will be the combination of the waiting times and the processing time at each and every step.

Lean Thinking versus Mean Thinking

Lean Thinking is a philosophy where waiting lists or inventory (business 'fat') are seen as the key symptom of waste in a System. Lean Thinkers recognise that the only cost effective, completely value-adding and efficient System is one in which there is no waiting for customers or staff. They recognise the impact of variation in Systems and systematically eliminate the causes of variation.

In contrast Mean Thinkers use the waiting lists, inventory or queue to buffer the utilisation of the resources against the variation in their System at the inconvenience of their patients, staff and customers. They strive to optimise the utilisation of their resources by using the queues as 'buffers' and fail to account for the hidden costs or waste incurred by the queues and waiting lists.

Summary.

Why do we get queues and waiting lists?

1. **Demand exceeds Capacity**

   If the average demand is greater than the average capacity then the queue or waiting list grows over time. This is rare in the NHS but an example of this is the waiting list or queue for follow-up appointments.

2. **The average demand = average capacity but there is a mismatch between the variations in demand and variations in capacity at each step.**

   This is the most common reason for waiting lists and queues in the NHS and explains why there will always be a queue beyond any target waiting time.

   This is because when the variations in demand and the variations in capacity mismatch, any demand that exceeds capacity will be carried forward as a queue, but when the capacity exceeds demand, the 'spare' capacity is lost. This means that there is always a deficit of capacity in the System. The backlog (queue) will wait until the next time the capacity exceeds the demand.

   In a System where the demand an capacity are varying, the average capacity required to keep a queue under control will need to be greater than the average demand. The amount of capacity required will depend on the amplitude and mismatch of the variations and the maximum waiting time that can be tolerated.

So to eliminate queues we have to:

1. Measure demand and the variation in demand.

2. Plan the average capacity at a level which allows for the variation in demand and the required response time. The more variation there is the more capacity is required to prevent a queue from building up.

3. Reduce the variation in the capacity (the main problem)
   a. Pooling according to the Pareto analysis of the case mix
   b. Level scheduling to smooth the capacity
   c. Standard work to eliminate error and rework

4. Reduce the variations in demand
   a. match capacity to meet predictable (special cause) variations in demand
   b. reduce the variations in upstream capacity.
   c. understand the demand for our service better: what do our patients actually want and when?

References:

There is a wide range of articles and books on ‘operations’ management but few, especially those specifically about the Lean Thinking, are taught on financial or business management courses.

References to Erlang and Pareto principle are found in any operations management text.

We have provided a reading list at www.steyn.org.uk and welcome any other references you may come across and have found useful.