gooroo for NHS

1

Waiting list management
An introduction

MAY 2010

Why are some hospital waiting lists harder to manage than others? How much difference do high urgency rates, removals, and patient cancellations make? Which booking tactics are most effective at coping with disruption? Is a fully-booked list more difficult to manage than a partially-booked one? When planning for the future, especially when waiting time targets must be met, how should we allow for all this?

We use simulation modelling to answer these questions, and the findings are presented in this suite of five papers. This first paper acts as an introduction, explaining and illustrating the simulation modelling.

gooroo®

What makes waiting lists so complex?

Waiting list management would be easy if every patient could simply wait their turn in the queue. It would be even easier if patients were always referred to the list at a constant rate. So long as the health service can provide enough capacity to keep up with demand, it can simply admit everybody on a first-come-first-served basis.

Unfortunately, real life is not so simple. The number of patients added to the waiting list varies randomly from week to week; some patients need to be seen more urgently than others; some patients are removed from the list before their admission date; and some patient appointments (and even entire sessions) are cancelled at short notice. All these events disrupt the orderly management of the waiting list.

In each of the following papers, we will use our simulation model to examine an aspect of waiting list management. The five papers look at:

1. Introduction: This paper, explaining the simulations.

2. How many urgent slots: A rule of thumb for the proportion of admission slots that should be reserved for urgent patients.

3. The causes of disruption: The effects of high levels of clinical urgency, big waiting lists, patient cancellations, removals, and fluctuations in referral rates.

4. Booking tactics: Good and bad ways of managing a waiting list, including full and partial booking, expediting urgent patients, rebooking long-waiting patients, allowing urgent patients to "bump" routine bookings, "rippling" bookings through the waiting list, and flexing capacity.

5. Waiting time targets: Establishes the link between the size of the waiting list and various kinds of waiting time target.

In all these papers we consider only persistent waiting lists, such as inpatient and outpatient waiting lists, where there are always significant numbers of patients waiting. Transient waiting lists, where sometimes there is nobody waiting (for example, patients waiting to be seen in an emergency department) are not covered here and are best understood using different mathematical approaches (notably queueing theory).

Simulating the waiting list

Computer simulation modelling is only useful if the simulation is realistic enough. For that reason we will explain below, in detail, how our simulation model works. A good computer simulation has substantial benefits compared with real-life experimentation:

1. There are no medical ethics issues to consider when experimenting with techniques that could risk delaying urgent patients;

2. The computer complies with its instructions exactly, so non-compliance with patient management rules is not an issue; and

3. It is possible to run large numbers of simulations to ensure statistical significance (the simulations run for these papers, for instance, are equivalent to studying several million doctor-years of clinical practice).

The main way in which the simulation model differs from real life is that our model provides a fixed number of slots per week, whereas in real life different patients require different amounts of time for their appointment. In real life this introduces the additional problem of scheduling patients to maximise the utilisation of the capacity available. Such complexity is beyond the scope of the present model, which assumes that a fixed number of appointment slots are available each week.

How the model runs

Each of our simulations runs through three phases:

1. initialisation: an initial waiting list is created, and the waiting list flux events for the model (additions, removals and cancellations) are set up to be random and yet conform to the scenario description;

2. settling: the model runs for a settling period of 50 weeks, without any measurements being made;

3. modelling: the model goes "live", measurements start being taken, and the model continues to run for ten consecutive periods of 50 weeks each.

For each week of the model, patients are:

a. added to the waiting list,

b. suspended or unsuspended from the list,

c. removed from the waiting list (without having their appointments),

d. booked in for their appointments,

and then, for all patients whose appointments fall due in the current week:

e. some or all have their appointments cancelled, and

f. the rest are admitted (and removed from the waiting list).

The process of booking patients for their appointments is also performed in stages:

i. unbooked urgent patients are given appointments;

ii. routine patients who have recently been cancelled are rebooked;

iii. routine patients whose booking will lead to an excessive wait are offered a rebooking;

iv. unbooked routine patients are given appointments.

The bulk of the simulations assume one clinical session per week, but the results are equally valid for two or more sessions per week (and indeed we look at this in some of the scenarios).

Demerit scoring system

Each scenario is evaluated by counting "demerit" points during the measurement period:

• delayed urgents: for each urgent patient admitted, who has waited longer than their "best-before" date:
2 demerits for each 1 week delay;

• long waits: for each routine patient admitted, who has waited longer than the ideal maximum waiting time (i.e. the maximum wait that would be achieved if all patients were admitted in line with clinical priorities without disruption):
1 demerit for every 2 weeks excess wait or part;

• rebookings: for each patient whose appointment is cancelled or moved for any reason:
1 demerit per rebooking; and

no-notice bookings: for each routine patient who is booked into an empty urgent slot in the current week:
1 demerit per no-notice booking.

The no-notice booking demerits are hardly visible in the quoted results, but they are important in the modelling of partially booked waiting lists. Without them, low demerit scores would be achieved by making all the appointment slots urgent and

only booking routine patients into the current week. The introduction of no-notice booking demerits penalises this practice, so that a sensible level of urgent slots can be discovered.

In all charts, demerit points are expressed as demerits per hundred patients admitted.

A simplified service

We will start by discussing a simplified service. In this example, patients are added to the list at a constant rate, are all "routine", are all given appointments on a first-come-first-served basis, the list remains a constant size, and there are no disruptions.

Intuitively, we would expect this service to be easy to manage: just give each patient the next available appointment slot, so that all patients are admitted in turn.

Simple service	
Patients added to the waiting list	10/week, every week.
Waiting list size	100 exactly
Patients needing an urgent appointment	None
Appointment slots available each week	10/week, every week
Patients permanently removed from the list before their appointment date	None
Individual patient appointments cancelled at short notice and rebooked	None
Whole sessions cancelled at short notice, and patients rebooked	None

Simplified service, fully booked

The following figure illustrates this simplified service, for the case where all patients have an appointment slot (i.e. a fully-booked service).

To explain:

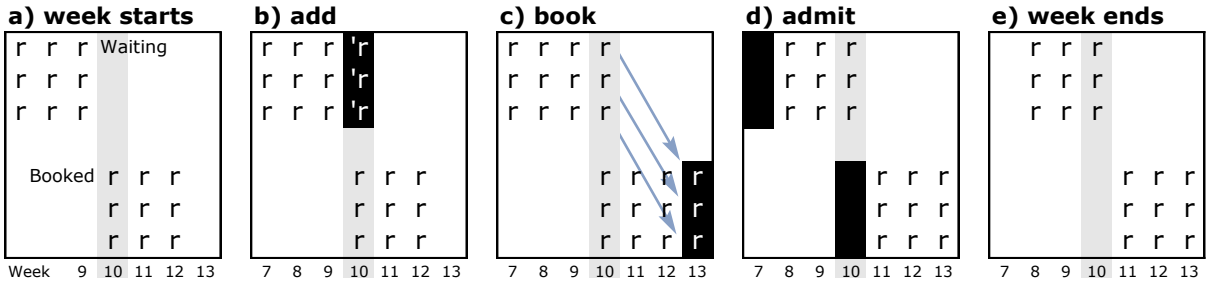


Figure 1: Simplified service, fully booked. See main text for an explanation.

Picture a) shows the position at the start of the current week (week 10, shaded grey). The top half shows the waiting list, and each "r" represents one routine patient. The waiting list is laid out according to the week when each patient was added to the list. Of the nine patients on the list this week, three were added in week 9, three in week 8, and three in week 7.

The bottom half of picture a) shows the appointments that have been booked for these patients. Of the nine patients on the waiting list, three have appointments this week, three next week (in week 11) and three the week after (in week 12).

Now the events of this week unfold.

Picture b) shows three more patients being added to the waiting list (highlighted in black). The tick in front of each letter "r" indicates that these patients do not have appointments booked yet.

Picture c) shows the new additions being booked for their appointments. Each patient is given the next available appointment slot, in week 13.

Picture d) shows patients being admitted. All patients with appointments this week are admitted (highlighted in black in the bottom half), and removed from the waiting list (highlighted in black in the top half).

Finally, picture e) shows the position at the end of the week. It is identical to the starting position, except that we are one week further on.

We guessed above that this simplified model would be easy to manage. The simulation has shown that our intuition was correct. If patients are simply given the next available appointment, they are all seen in turn, and maximum waiting times are perfectly controlled. The number of demerit points in this example is zero.

Simplified service, partially booked

There is another way to manage this waiting list: partial booking. Instead of giving appointments to all patients on the list, we could only issue appointments (say) two weeks in advance. Then the simulation would look like this:

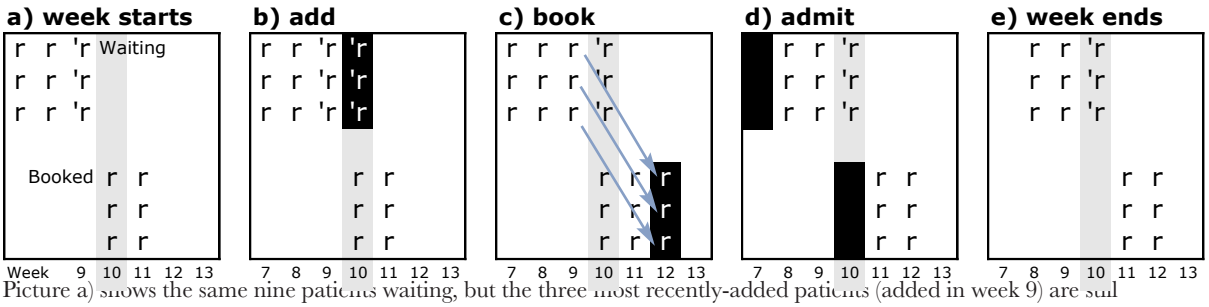


Figure 2: Simplified service, partially booked. See main text for an explanation.

Picture a) shows the same nine patients waiting, but the three most recently-added patients (added in week 9) are still unbooked. Only six bookings are shown in the bottom half of the picture.

Three more patients are added in picture b), also unbooked. (Remember that the tick before the "r" means the patient is unbooked.)

In picture c), three appointment slots have been opened for bookings in week 12, and the longest-waiting unbooked patients are booked into them.

In picture d), patients with appointments this week are admitted.

Picture e) shows that the position at the end of the week is the same as at the start.

So the simulation shows that this service is equally simple under partial booking. A first-come-first-served regime results in perfect list management and no demerit points.

Simplified service, illustrating whole session cancellation

Our final example of a simplified service shows the effect of cancelling an entire session.

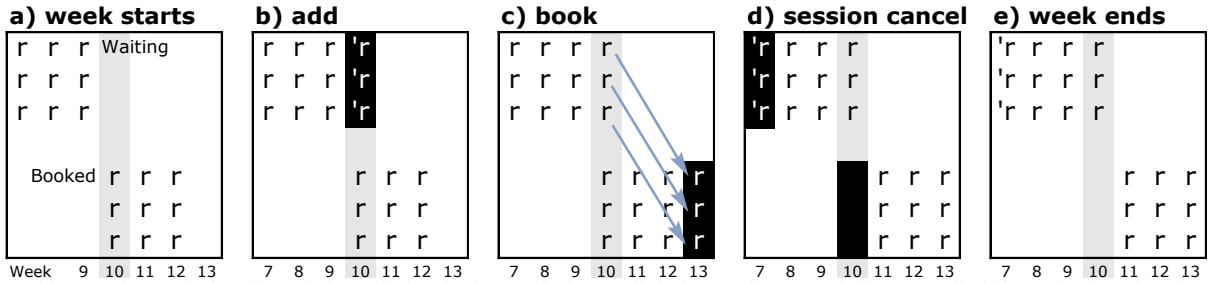


Figure 3: Whole session cancelled, fully booked regime

In this case, instead of patients being admitted, picture d) shows all their appointments being cancelled (from the bookings in the lower half), and the patients remaining on the waiting list (top half) without bookings. Because this is a fully-booked regime, they will all be booked into the next available slots (in week 14) during the coming week.

Adding complexity

Urgent patients and other real-life disruptions

Unfortunately, real life is not as simple as the scenario above. The following figure shows a more complex waiting list including urgent patients and variable numbers of patients being added from week to week.

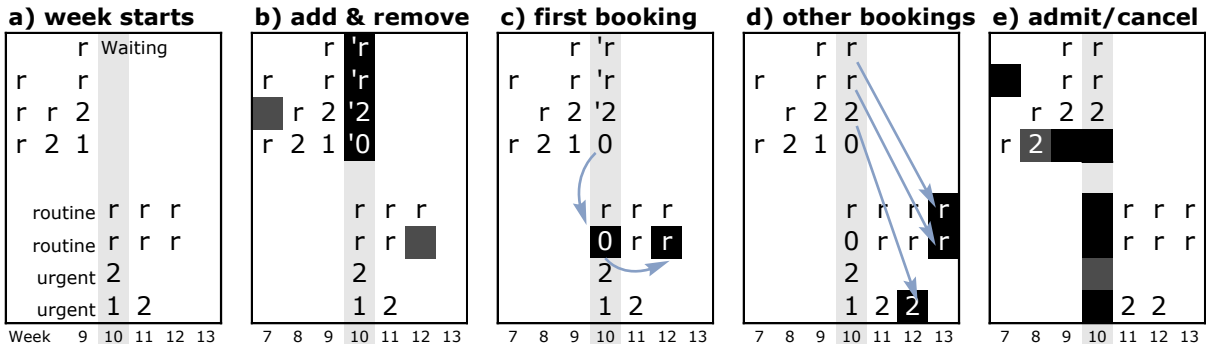


Figure 4: More realistic service. See main text for an explanation.

Picture a) shows the position at the start of the week. The top half shows that there are nine patients on the waiting list: six routine patients and three urgents. Each urgent patient is represented by a number showing how long they can safely wait from the time they are added. For instance, the urgent patient added in week 8 needed admitting within 2 weeks, and must be admitted by week 10 at the latest.

The bottom half of picture a) shows the appointments booked for these nine patients. Four slots are available each week, and this capacity has been partitioned into two urgent slots and two routine slots. This partition protects capacity for urgent patients arriving at short notice. The current week (week 10, highlighted in grey) has appointments for two routine patients, the 2-week urgent patient added in week 8, and the 1-week urgent patient added in week 9.

In picture b), another four patients are added to the waiting list (highlighted in black): two routines, a 2-week urgent (who must be admitted in week 12 at the latest) and a 0-week urgent (who must be admitted this week). One patient is also removed from the waiting list (highlighted in dark grey): a routine patient who was added in week 7 and booked for an appointment in week 12.

Suspension is not shown here (it is not part of the reference scenario, but the effects are examined in later papers). When patients are suspended they are unbooked and removed from the waiting list. Later on, when they are unsuspended, they reappear on the waiting list with their original waiting time preserved (and will later be rebooked). So if a patient who was added in week 8 is suspended in week 10 (having waited 2 weeks), then if they are unsuspended in week 20 they will reappear as if they had been added in week 18 (having waited 2 weeks).

We have a problem with the 0-week urgent patient, because there are no empty appointment slots this week. Picture c) shows a possible solution: "bumping". We cancel a routine patient's appointment, and replace them with the urgent patient. Then we rebook the displaced routine patient into the next available routine slot (which happens to be the slot vacated by the patient removed in picture b).

Now we can book the remain patients, as shown in picture d). The unbooked 2-week urgent patient is booked into week 12, even though an earlier urgent slot exists in week 11; this practice is referred to as "searching backwards for an empty urgent slot". Then the two unbooked routine patients are given the next available routine slots.

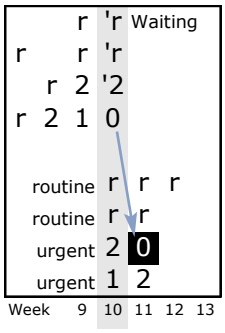
Finally, picture e) shows this week's appointments session being held. One patient (a 2-week urgent, highlighted in dark grey) is cancelled and remains on the waiting list. The other three are admitted and removed from the list.

The end-of-week picture is not shown here, but we can see that it will not be identical to the start-of-week picture. This is usually the case in a real-life waiting list with all its disruptions.

Alternative booking tactics

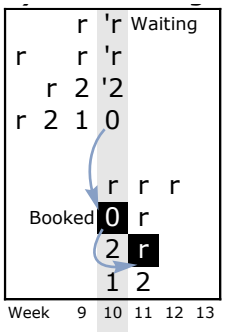
Here are four alternatives to picture 3c) above, in which we had to book a 0-week urgent patient but had no empty slots available in the current week.

Figure 5: Forbid bumping



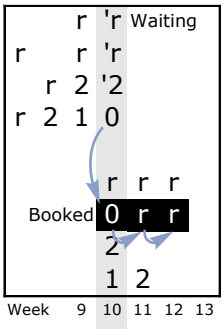
If "bumping" is forbidden, then we cannot book the recently-added 0-week urgent patient into week 10 by displacing a routine patient. Instead we must give them the next available urgent or routine slot: in this case the empty urgent slot in week 11. This imposes a 1-week delay on this urgent patient and will attract 2 demerit points.

Figure 6: Allow bumped or cancelled routines to use urgent slots



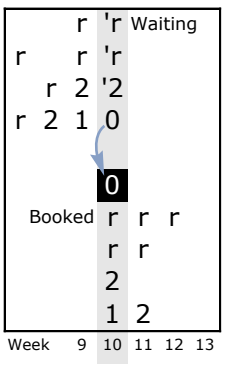
In this picture, bumping is allowed (and we can book the 0-week urgent on time), but we rebook the displaced routine patient in a different way. Instead of giving them the next available routine slot, which might cause a significant delay, we allow them to book into an earlier urgent slot. In this case they book into week 11, and only suffer a 1-week delay as a result of being bumped. In the simulation model, we can specify a minimum delay (of, say, 4 or 8 weeks) to reduce the disruption to urgent patients.

Figure 7: Rippling



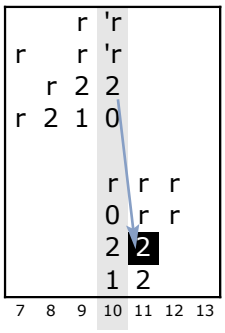
Rippling is another way to manage the displaced routine patient. If we do not allow them to use an urgent slot, then we can minimise their delay by allowing them to bump another routine patient. In this picture, the urgent patient bumps a routine patient from their week 10 appointment; the displaced routine patient bumps another from their week 11 appointment; and this last patient is given an empty appointment slot in week 12. So each displaced patient suffers a delay of only 1 week, but two patients are displaced instead of one. In the simulation model we can specify the maximum delay per patient (eg. 4 weeks), and the model bumps the patient within that limit who will experience the shortest waiting time.

Figure 8: Flexing capacity



Another way of accommodating the 0-week urgent patient is to squeeze them in by extending the current week's appointments. This is illustrated in the picture. There is a fair amount of complexity around doing this in practice, and we will discuss this in more detail below.

Figure 9: Search forwards for urgent slot

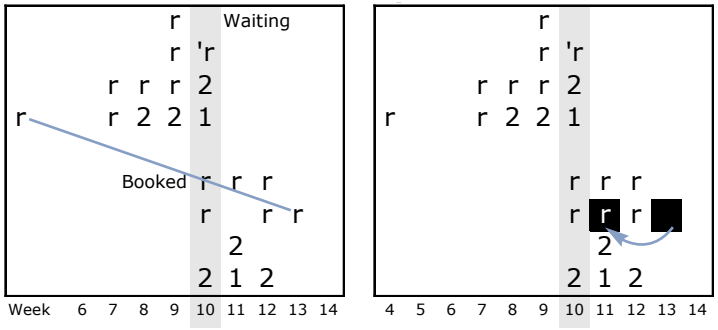


This is an alternative to the booking of a 2-week urgent patient in picture 3d), when we started at the best-before date (week 12), searched backwards for an empty urgent or routine slot, and booked the patient into an empty urgent slot in week 12. In this alternative picture we start in the current week (week 10) and search forwards for an empty urgent or routine slot; the first empty slot is in week 11 and the patient is booked into it, even though the patient could safely wait a week longer.

Rebooking potential long waits

The final tactic we will look at is rebooking potential long-waiting patients.

Figure 10: Rebooking a potential long-wait into a future empty routine slot



In the first picture of Figure 10, the top half shows that the waiting list includes one patient who has been waiting since week 4. A line links this patient with their appointment in week 13: a potential wait of 9 weeks. This is much longer than the 3-4 week wait being experienced by most other routine patients on this list.

The second picture shows how this patient could be rebooked into the empty routine slot in week 11, shortening their wait by 2 weeks. This would reduce the long-wait demerit score by 1, but attract a penalty of 1 demerit score for the rebooking. In the model we can set a minimum reduction in waiting time before rebooking is allowed, to ensure that a net improvement in score is always achieved.

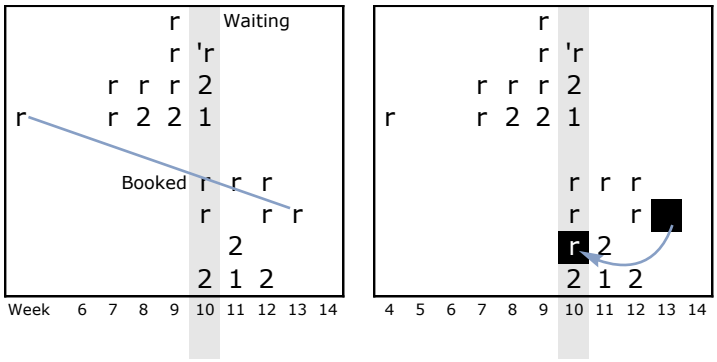


Figure 11: Rebooking a potential long-wait into the current week

Figure 11 shows a different approach to rebooking: using any empty slots that remain in the current week, whether urgent or routine.

For both types of rebooking, in the simulation model we can change the likelihood that a potential long-waiter will accept a rebooking. We can also change the definition of "potential long-waiter" relative to the ideal maximum waiting time, to determine which patients are offered rebooking.

More about flexing capacity

Flexing capacity involves making each session longer or shorter in response to the immediate pressures of the current week.

We may want to run a longer session because of:

• urgent patients who cannot find slots within their best-before dates (by creating extra capacity for them, we remove the need to bump a routine booking);

• routine patients whose current appointment will result in an excessively long waiting time; and

• patients whose appointments have just been cancelled at short notice.

Those are all reasons for making this week's session longer. But we need to balance this with reasons for making this week's session shorter. (Otherwise all we will achieve is the creation of extra capacity and the evaporation of the waiting list; not a very interesting modelling result.)

We have an opportunity to run a shorter session when:

• individual patients are cancelled at short notice (for instance, because they failed to attend their appointments); or

• whenever there are empty slots remaining in the current week.

Finally, we need a way to ensure that longer and shorter sessions roughly balance each other out. Usually longer sessions will, on average, outweigh the opportunities to run shorter ones. So when the longer sessions exceed the shorter ones by (say) 5 appointment slots, we impose a moratorium on longer sessions, until we have accumulated enough shorter sessions that the balance is (in this example) 5 slots in favour of the shorter sessions. Then we allow longer sessions again.

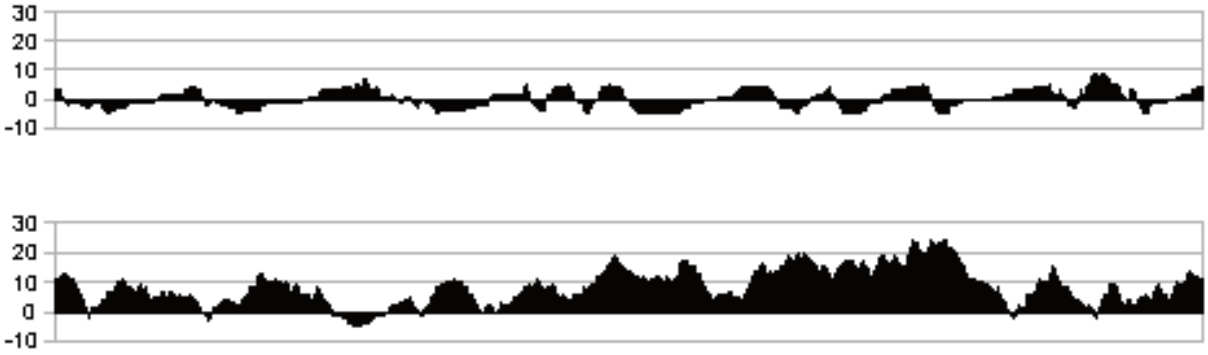


Figure 12: Accumulated flexibility over a real 500-week modelling run.

Top chart is for 39% urgent slots, bottom chart is 41% (the optimum).

The above chart shows an example of the accumulated flexibility over time. It turns out that the optimum demerit score is achieved when the booking rules are on the point of achieving a long-term surplus of vacant urgent slots (meaning that the moratorium on longer sessions is rarely imposed and flexibility is nearly always available). With the booking rules set just a little lower, as in the top chart, the moratorium on flexibility is imposed more often.

It is worth saying a few words about some other methods of flexing capacity which were tried, but failed, because they would also come to grief if they were tried in real life.

1) It might seem logical to run a longer session in response to this week's pressures, and repay the time by planning a shorter session some time into the future (just beyond the booking horizon for routine patients, for instance). This fails because the pressure to run longer sessions outweighs the opportunities to run shorter ones, and so the debt is perpetually being extended and never repaid.

2) We tried removing any cap on the outstanding time debt, and trying to choose booking rules that achieved balance in the long run. This did not work because the system is too sensitive: runs of longer or shorter sessions (which would be difficult to sustain in real life without complaints from staff) cause large fluctuations in the size of the waiting list.

3) Finally, we tried using flexibility rules which continuously adapted to the outstanding time debt. This created a time-lagged feedback loop resulting in increasingly wild fluctuations, as the rules over-corrected for the deviations established previously.

An extract from the real simulation model

The examples above are simplified illustrations in order to explain particular points. The real simulation model is bigger and has to handle more complexity. Fortunately, computers are good at handling this. Here is a sample snapshot:

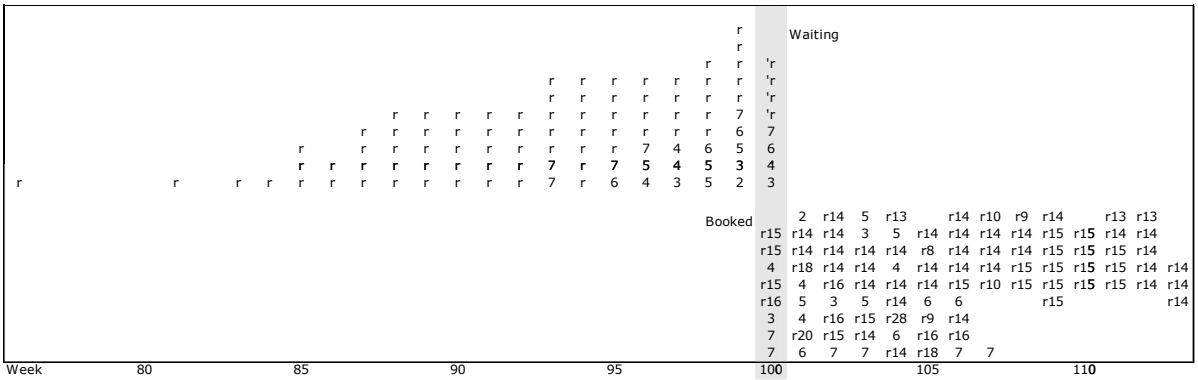


Figure 13: Snapshot from the simulation model, showing the waiting list (top half) and appointments booked (bottom half)

This snapshot is taken in week 100, just after the newly-added urgent patients have been booked, but before the routines are booked. It is based on the reference scenario specified below in this paper.

The notation is the same as in the simplified examples above, except for one additional piece of information: in the booking section (bottom half) each routine patient is tagged with their potential waiting time (i.e. how long they will have waited when their appointment falls due). For instance, the longest-waiting patient at the extreme left of the waiting list (left of top half), who was added to the list in week 76, has been given the 28-week routine booking in week 104 (marked "r28").

The reference scenario

The results in this suite of papers are presented relative to a reference scenario. The reference scenario has been chosen to combine typical levels of waiting list disruption with good booking tactics, as follows.

Reference scenario: Attributes and causes of disruption

Patients added to the waiting list	Average 10/week. Varies ±4 from week to week (binomial distribution between 6 and 14).
Target average waiting list size	100 patients on average. It varies randomly from week to week, but this is balanced to ensure that the start and end list size is as close to 100 as possible, and so is the average list size over each 50-week period.
Patients needing an urgent appointment	40% urgent patients, needing admission within 8 weeks. The profile of urgencies is evenly spread, so 5% need admission within 1 week, 5% in 1-2 weeks, and so on up to 7-8 weeks. Urgencies are allocated randomly to each patient according to this overall profile.
Appointment slots available each week	8.8, calculated to achieve the target average list size, and equals (average additions per week) * (1 - % removed + % patient and session cancellations)
Suspensions	None
Patients permanently removed from the list before their appointment date	15% (close to the England average for inpatients and daycases). Each patient on the list has an equal appointment date chance of being removed each week.
Individual patient appointments cancelled at short notice and subsequently rebooked	2%, randomly selected from those patients whose appointments fall due each week (roughly the England average for inpatient and daycase Did Not Attend rates).
Whole sessions cancelled at short notice, and patients subsequently rebooked	1%, randomly triggered as each weekly session falls due (which is roughly the England average for inpatient and daycase cancellations for non-clinical reasons).

Reference scenario: Booking tactics

Slots reserved for urgent patients proportion is the subject of the second paper in this series.)	43% of slots are reserved for urgent patients; the remaining slots are 'routine'. (The optimum
Urgents always booked earlier than their best-before dates?	No, a safety margin is not built in.
How to search for empty slots for urgent patients?	Start at best-before date and work backwards, so that urgent patients wait as long as possible within clinical safety.
If no empty slots are available before the best-before date, allow urgent patients to displace ("bump") booked routine patients?	Yes, allow bumping
Flex capacity by varying the length of each session in response to short-term pressures?	No, do not flex capacity
Allow rippling?	No, do not ripple
When searching for the next available empty slot for a bumped or cancelled routine patient, allow them to take an empty urgent slot?	Yes, allow displaced or cancelled routines to use urgent slots, if no earlier routine slot is available, with a minimum delay of 4 weeks.
How to book other routine patients	Use next available empty routine slot.
If a slot remains empty in the current week, should a potential long-waiting routine patient be rebooked into it?	Yes, try to rebook patients who will wait longer than the ideal maximum wait; assume a 40% probability that they will accept the rebooking.
If a routine slot remains empty in future weeks, should a potential long-wait be rebooked into it?	No.
Give appointments to all patients on the list?	Yes, run a fully-booked regime

Index of figures

The four further papers in this series will evaluate the situations and tactics illustrated here. So we will finish this paper with an index of the pictures which best illustrate them:

Attributes and causes of disruption

Patients added to the waiting list	Constant rate: Figures 1, 2 and 3 Variable rate: Figure 4
Waiting list size	Constant: Figures 1 and 2 Variable: Figures 3 and 4
Patients needing an urgent appointment	No urgents: Figures 1, 2 and 3 Some urgents: Figure 4
Patients permanently removed from the list before their appointment date	None: Figures 1, 2 and 3 Some: Figure 4
Individual patient appointments cancelled at short notice	None: Figures 1, 2 and 3 Some: Figure 4
Whole sessions cancelled at short notice	None: Figures 1, 2 and 4 Some: Figure 3

Booking tactics

Slots reserved for urgent patients	None: Figures 1, 2 and 3 Some: Figure 4
Urgents always booked earlier than their best-before dates?	No expediting: Figure 4 Some expediting: not illustrated, but if urgents are being expedited by 1 week, then a "2-week urgent" would be booked as a "1-week urgent".
How to search for empty slots for urgent patients?	Search backwards: Figure 4d Search forwards: Figure 9
If no empty slots are available before the best-before date, allow urgent patients to displace ("bump") booked routine patients?	Allow bumping: Figure 4c Forbid bumping: Figure 5
Flex capacity by varying the length of each session in response to short-term pressures?	No flexing: Figure 4c Allow flexing: Figure 8
Allow rippling?	No rippling: Figure 4c Ripple: Figure 7
When searching for the next available empty slot for a bumped or cancelled routine patient, allow them to take an empty urgent slot?	Allow: Figure 6 Forbid: Figure 4c
Should a long-waiting routine patient be rebooked into an earlier empty slot?	Into empty routine slot in a future week: Figure 10 Into empty urgent or routine slot in current week: Figure 11
Give appointments to all patients on the list?	Fully booked: Figures 1, 3 and 4 Partially booked: Figure 2

CONTACT US

BY POST
Gooroo Ltd
The Old Grammar School House
School Gardens
Shrewsbury, Shropshire
SY1 2AJ

BY PHONE
01743 232149

BY EMAIL
info@gooroo.co.uk

ONLINE
gooroo.co.uk