An edited version of this article appeared as: Jones R (2001) Waiting times: quick, quick, slow. Health Service Journal 111(5778), 20-23. Please use this as the citation.

Guaranteed Urgent Appointments?

Dr Rodney Jones (ACMA)
Statistical Advisor
Healthcare Analysis & Forecasting
www.hcaf.biz
hcaf_rod@yahoo.co.uk

Key Points

- Guaranteed waiting times require over-provision of resources in order to cope with very high randomness in weekly arriving referrals
- Randomness makes it virtually impossible to allocate the 'correct' number of urgent appointment slots
- Current appointment systems are not able to cope with the operational demands arising out of randomness
- Achievement of the 2 week cancer referral waiting time will be subject to the forces of randomness
- Randomness dictates that only a proportion of appointments can be offered a guaranteed time

Introduction

Recent focus on achieving waiting time targets for outpatients and inpatients has led to increased awareness of the role which allocation of urgent outpatient or theatre slots plays in influencing the waiting time of both the urgent and non-urgent patients. The promised maximum two week wait for cancer referral has also led to a perceived need for some of the urgent outpatient slots in a particular specialty to be reserved for cancer patients.

The issue appears to be almost trivial. If the urgent waiting time is too high simply increase the relative allocation of urgent slots, either for a first outpatient appointment or for an urgent operation, and thereby the problem is solved. While equally relevant to outpatient and inpatient waiting lists the following discussion will use an outpatient context to illustrate the complexity of the issues involved.

Discussion with Consultants reveals that the assumed simplicity of appointment slot allocation is highly misleading. The following examples are relevant:

- In some specialties there are significant numbers of urgent follow-up patients whose condition can flare up at any time (e.g. Rheumatology, Cardiology, Gastroenterology, Mental Health, etc). These patients become part of the total demand for urgent outpatient services.
- Consultants make adaptive responses to urgent needs and will see patients at very short notice in a follow-up clinic where it is hoped that a DNA will create room for the urgent patient. It is far easier to do this for the occasional patient and where the volume of urgent patients is low.
- The mix of patients classified as urgent varies between specialties. For example, in ENT the bulk of urgent patients will be suspected cancers, however in other specialties the proportion can be very small.
- The classification of urgent is not clear cut and ranges from the equivalent to an emergency outpatient attendance through to a ‘soon’ patient who can be seen earlier if urgent outpatient waiting times are low at the moment.

The end result is that considerable flexibility is required which is incompatible with simplistic allocation of appointment slots. However, most booking systems dictate that there must be a clear-cut allocation of appointments. How else can patients be given the fixed appointment date and time stipulated by the Patients Charter and more recent booked appointment directives?

Some estimate of urgent, non-urgent and follow-up volume needs to be made and appointment slots allocated accordingly. However, if too many urgent slots are allocated the waiting time of the non-urgent patients will increase and vice versa.

Most managers would agree that there must be a better way than trial and error to achieve these objectives. Particularly so because guaranteed waiting time targets have the implicit assumption that a guaranteed time is deliverable in practice. Applying statistical process control to routine waiting times has led to this assumption being questioned (1).

Fortunately a particular type of statistics called Poisson probability allows us to develop our understanding of such problems (2). Poisson statistics is the basis for what is called queuing theory, which gives a description of the dynamic behaviour of any queue. To achieve this two fundamental pieces of information are required, namely, the arrival rate (e.g. number of urgent referrals per week) and the service rate (e.g. number of new appointment slots available each week to the arriving patients less DNA patients).

The apparent simplicity of allocating, say, 2 slots per week to an expected 2 arrivals per week is in fact shattered by the fact that Poisson statistics tells us that outcomes other than the average are highly likely. In fact a Poisson distribution changes shape depending on the expected average and becomes highly skewed as the expected average decreases. In order to make the following discussion of practical relevance we must first establish the level of urgent referrals received by most consultants.

A review of 52 outpatient clinics at a large DGH (a subset of the total number of clinics) showed that only 5 Consultant clinics (10%) had more than 4 urgent slots per week. The proportion of first appointment slots allocated to urgent appointments had a mode of 33% (minimum value 5%), however, just over half of the clinics allocated fewer than 25% of first appointment slots as urgent. Those with the lower proportion allocated to urgent appointments tended to be clinics with a medical emphasis although Cardiology is a notable exception with 65% urgent appointments. The fact that the most common value was 33% seems to indicate that a default value of 1 in 3 has been selected by Consultants as a preferred method to allocate urgent slots.

In the context of cancer referrals the largest weekly average is for combined upper and lower GI cancers where a typical large DGH would receive around 10 to 20 per week. This is spread over a number of consultants and hence the average per consultant will be less than 10 per week. In fact in most instances the highest average of new cancer referrals per consultant is usually less than 5 per week.

As discussed above the situation for the general category of urgent first appointments gives low average arrival rates for all consultants with almost all consultants receiving fewer than 20 urgent requests per week and the majority receiving fewer than 10 per week. The highest possible level of urgent appointments is probably to a rapid diagnosis breast clinic where a single consultant firm could receive up to 2,000 new referrals per annum. If we assume all these referrals are urgent (and they are not) this gives a maximum possible urgent demand of 40 per week.

Having established the boundaries we can now specifically investigate the effect of Poisson randomness on such small number events. This is summarised in Table One. Several important points emerge from a consideration of Table One, namely:

- Even at an expected arrival rate of 20 per week it is possible (although with very low probability) to get one week in which there are no arrivals
- The average arrival rate does not occur with high likelihood
- There are a higher proportion of weeks when there are less arrivals than the expected average (which can lead to the mistaken impression that low waiting times are always achievable)
- The skew to values lower than the average increases as the average reduces, e.g. at an average of one arrival per week it is 30% more likely to get no arrivals than any number higher than one, i.e. 37% compared to 26%

Table One: Likelihood of different outcomes given an expected average arrival rate per week.

Average urgent arrivals per week	% of weeks with no arrivals	% of weeks with fewer arrivals than the average	% of weeks with more arrivals than the average	% of weeks with average number of arrivals
1	37%	37%	26%	37%
2	14%	40%	32%	28%
3	5%	42%	35%	23%
4	2%	43%	37%	20%
5	1%	44%	38%	18%
6	0.2%	45%	39%	16%
7	0.1%	45%	40%	15%
8	0.03%	45%	41%	14%
9	0.01%	46%	42%	12%
10	0.005%	46%	44%	10%
20	0.0000002%	47%	44%	9%
40	4 x 10^{-16}%	48%	46%	6%

These observations lead us to a further uncomfortable question. How do we actually know the true average expected arrival rate? The majority would respond by saying that we count the referrals and take an average. Managers who have studied statistics would also point out that most textbooks indicate that it takes around 30 to 40 measurements to establish an accurate average. This would imply that if we measure the arrivals for 30 to 40 weeks and take an average we should have an 'accurate' measure of the true average. In practice seasonal effects on referral rates and the occurrence of public holidays make anything less than a 52-week sample subject to considerable bias (3). However, Poisson statistics does have particular requirements that are not usually discussed in general statistical textbooks. Table Two shows the accuracy obtained from one, two and three year sample periods.

Table Two clearly shows that the accuracy of any attempt to estimate the average declines rapidly as the average arrival rate declines below 20 per week, i.e. the case for almost all outpatient clinics. It also clearly shows that a three year sample period is required to gain any reasonable level of accuracy – assuming that referral rates are constant over this period. This observation goes a long way to explaining why clinicians have such difficulty in determining the correct allocation of urgent slots – it is an almost impossible task. So don't blame the clinicians for this one!

Table Two: Effect of the number of measurements on the accuracy of the calculated average

True average arrivals per week	Maximum uncertainty in the calculated average given different sample sizes		
	52 weeks	104 weeks	156 weeks
1	50%	30%	26%
10	19%	11%	8%
20	9.5%	7.5%	5%
40	7.5%	5.5%	3.8%

The combined results of Tables One and Two clearly reveal the extent of the gulf between clinical needs and administrative requirements. All outpatient clinics experience high randomness in the number of urgent referrals received in any week. The exact number arriving cannot be predicted and hence for the purpose of allocating appointments we have to make a planning assumption. Even if we choose to plan at the average the randomness underlying the referrals means that our estimate of the average is likely to be high or low. Are there any solutions to this apparently insoluble conundrum?

A study of Poisson statistics shows that the standard deviation associated with the average is described by the square root of the average. To those who are not statisticians, the standard deviation is simply a measure of how widely the results are scattered around the average. High standard deviation implies high scatter and hence by implication a low possibility of being at the average. As a good approximation we can also say that the maximum and minimum possible values are equal to the average plus or minus three times the standard deviation. This is not strictly true for a Poisson distribution but it is a reasonable approximation (2).

Hence for an expected 9 referrals per week the standard deviation will be three times the square root of 9 which is 3, hence 3 x 3 = 9. We can therefore expect between 9 ± 9 or between 0 and 18 referrals in any week.

Poisson randomness applies to the whole year as well as the individual weekly results. Hence if we are expecting 9 urgent referrals per week (range 0 to 18) over a full year we are expecting 9 x 52 = 468 new referrals. Hence our maximum and minimum possible referrals over the year becomes 468 ± 65, i.e. somewhere between 403 and 533 new referrals. This implies that if we plan to see 468 patients then by year end our number waiting to be seen can have increased or decreased by as many as 65. The effect of this on the waiting time is fairly obvious and is discussed later.

Given our observation that the largest volume in terms of urgent referrals will be to a rapid diagnosis breast clinic this implies that the best possible performance attainable will be 2,000 ± 134. This implies (ignoring the effect of DNA's), that to avoid an increase in waiting time the breast clinic would have to plan for 2,134 new appointments per annum. This represents a 7% over-provision against the expected average.

This over-provision will increase with decreasing volume such that at an average of 1 urgent referral per week the number of clinic slots required to avoid any increase in waiting time will be 40% more than the average, i.e. 73 as opposed to 52 appointment slots per annum. Should referrals be low due to randomness then it is possible that up to 58% of clinic slots would be empty. This represents a very high potential wastage of scarce resources arising from the consequences of Poisson randomness.

Turning to the question of guaranteed maximum waiting times Poisson statistics can also shed valuable insight into whether such targets are achievable (i.e. can be guaranteed under all circumstances) and to what extent they could lead to potential wasteful over-provision.

For instance, the guaranteed two week waiting time for urgent cancer referral can be shown to be either operationally impossible or lead to massive over-provision of appointment slots resulting in extended waiting times for non-cancer referrals. If we make the assumption of 10 urgent cancer referrals per week then we have the possibility of 520 ± 68 referrals per annum, i.e. between 452 and 588 per annum.

To guarantee a maximum wait of two weeks implies the provision of 588 appointment slots per annum with a maximum possible (worst case) over-provision of 166 appointment slots. This implies that up to 166 non-cancer urgent appointments may have to wait longer because of the selective allocation of scarce resources. This is of course the worst case scenario and the actual outcomes across all hospitals will range between the two extremes.

To put this in context it is worth noting that the April to June national figure of 92.4% of cancer patients seen within two weeks is consistent with the skew in a Poisson distribution. Actually taking this figure higher will be confounded by the tail of high arrivals implicit in a Poisson distribution.

However, the obvious implication is that unless it absolutely cannot be avoided there is never any justification for the provision of cancer-only appointment slots within a specialty. It is simply too wasteful of scarce resources. The ideal solution is to pool cancer and non-cancer urgent appointments to gain the benefits of increased size and so reduce the over-provision required to guarantee a maximum wait.

Having concluded that there is no basis for separating cancer and non-cancer referrals we need to look at the maximum change in the waiting time which can arise from a mismatch between referrals and appointment slots.

Maximum increase in waiting time

Before discussing the various alternatives which can be used to attempt to address the difficulties in allocating the correct number of urgent appointment slots

The maximum increase in the waiting time can be calculated using the following formula:

$$\text{Maximum increase in waiting time (weeks)} = \frac{N \text{ x } (R - S) + 3 \text{ x } \sqrt{(N\text{x}R)}}{R}$$

Where,

N = number of weeks in the period
R = average urgent referrals per week (in the period)
S = number of slots provided per week (in the period)

This formula is useful because it allows for seasonal change in referral rates and periods when the number of appointment slots are increased or reduced for any reason. It can also be used to calculate the number of slots required to reduce the waiting time to an acceptable level.

Note that the formula gives the increase in the waiting time for a period of 'N' weeks from the current date and hence the maximum waiting time will be equal to the current waiting time plus the potential increase in waiting time. To locate the maximum weekly variation in waiting time simply put N equal to 1 week.

Having established the maximum possible variation in waiting time resulting from randomness in the arrival of referrals we can now progress to a review of the options available for the allocation of appointment slots.

Most of the following options are based on the use of the annual rather than the weekly average number of referrals. We are forced to do this to avoid the high variation that occurs at the weekly level. However, to quantify the full impact of randomness we must check the resulting conclusions against the weekly level of randomness. The various options available are as follows:

Minimum provision with option for extra as required

Using this option each clinic would provide the minimum number of expected urgent appointment slots for a whole year. Rather than use three times the standard deviation it is probably best to provide two standard deviations below the expected average. This is to take account of the uncertainty in the average (as in Table Two) and the fact that a Poisson distribution has a longer tail at values higher than the average. In fact in a Poisson distribution only 2% of all possible outcomes are less than two standard deviations below the average, hence, 98% of all possible outcomes will be higher than this value.

For example, if we were expecting 100 urgent referrals per annum we would provide a minimum of 80 urgent appointment slots per annum. In practice this minimum would have to be adjusted upward to account for the higher than 3.5% DNA rate associated with urgent appointments (4).

At the same time we would allocate a series of reserve clinics containing five times the standard deviation as the number of reserve slots available to cope with random variation. For an average of 100 urgent referrals per annum this would imply the provision of 80 fixed slots and 50 reserve slots, i.e. 8 fixed slots every five weeks and 1 reserve slot per week or perhaps 2 at the end of each fortnight. The reserve slots could be run as the equivalent to an over-booked clinic or in a special clinic depending on the availability and flexibility of resources.

This option has the advantage that there is no wastage of scarce resources. Its limitation is the need for flexibility or overbooking on particular occasions.

It also has the limitation that the waiting time for an urgent appointment can vary considerably. For instance, an average of 100 referrals per annum is approximately 2 per week and this allows a maximum of 7 referrals (0.3% probability) in a single week. Should 7 referrals arrive in one week then the next arriving referral would have a wait of 3 weeks longer than the current waiting time.

Hence this approach to allocating appointment slots is resource efficient but not suitable to those specialties where the waiting time can not vary considerably for whatever reason, i.e. cancer waiting time. It is however an appropriate strategy for a specialty able to cope with some variation in urgent waiting time.

Provision at the expected average

Within this option we simply determine the expected average of demand (within the limitations of Table Two) and accept that there will be some over-allocation and hence wastage of resources. We also know that we will need to provide for up to three times the standard deviation for those occasions when demand is higher than the average. Hence, for 100 urgent referrals per annum our maximum reserve provision would be 30 per annum or 3 slots every 5 weeks.

This method has the advantage of only a moderate wastage of resources, e.g. maximum possible wasted appointment slots will be three times the square root of the expected average. Potential for wastage will obviously increase as the volume reduces, e.g. maximum wastage of 6.7% at 2,000 referrals per annum and 21.3% at 200 referrals per annum. It still has the disadvantage that there is potential for increase in waiting time as per the above equation.

Provision at the expected annual maximum

This would be the alternative of choice if there were need for a near absolute guaranteed waiting time or where waiting time variation needed to be minimised. As with the option for minimum provision it may not be necessary to provide a total number of appointment slots of the average plus three times the standard deviation (e.g. square root of the average). For any volume of referrals above 100 per annum it is sufficient to provide 2.4 standard deviations above the average. This will cater for 99% of all possible outcomes.

Since we are using annual totals to calculate the weekly number of appointment slots there is still the possibility of some increase in the waiting time due to weekly randomness as per the above equation. This method has the disadvantage of much higher levels of wasted appointments, e.g. up to 5.4 times the square root of the average. Hence at 2000 referrals per annum there is the potential to waste up to 12% of appointment slots and at 200 per annum the potential wastage is 38%.

Full Provision based on expected weekly maximum

This alternative would be chosen if there were need to give an absolute guaranteed waiting time such as a 2 week maximum wait for an urgent cancer referral.

As in the option above the full provision of average plus 2.4 times the square root of the average weekly volume would be sufficient to cover 99% of all possible outcomes. However, we are now seeking to provide clinic slots based on weekly averages of incoming referrals. This means that we are taking the square root of small numbers (in all cases less than 40 and in the majority of cases less than 5). As expected this method of resource allocation is highly wasteful with a minimum wastage of 40% for the largest of clinics (2,000 urgent appointments per annum).

This is the inevitable outcome of a guaranteed waiting time and suggests that most hospitals will not have the surplus resources required to meet the national guaranteed cancer waiting time target. If they do allocate sufficient appointments to guarantee the target the waiting time for non-urgent appointments will suffer greatly.

Clinical vs managerial requirements

Consultants are intelligent people and will attempt to allocate resources in a flexible manner – as much as is possible within the constraints of fixed booking systems. The discussion above referred to the potential wastage of scarce resources. In practice the simple throughput of patients is only part of the role of a Consultant. Their role in teaching implies that at those times when there is a full clinic they would do less teaching/supervision and at those times when there are 'empty' slots they would tend to do more teaching/supervision.

Hence from a consultant's point of view the scarce resource is not wasted but utilised in a different manner. However from a throughput point of view the opportunity to see an extra patient has been wasted.

The aim then should be to provide opportunities for teaching while not wasting too many opportunities to see an extra patient. This would tend to indicate that the 'best' solution is for perhaps the expected average plus one standard deviation. Around 90% of all expected outcomes would be covered leaving the 10% of remaining outcomes to be covered via occasional overbooking of urgent patients into follow-up slots or similar strategies.

Conclusions

The allocation of urgent appointment slots does represent a planning dilemma. The above discussion clearly shows why this has been such a difficult area for the planning of outpatient services. The volume of urgent appointments is simply too small to prevent high inherent randomness. This randomness is beyond the control of any health service body. As a result almost all the options available lead to the potential wastage of scarce resources and will result in an increase in the waiting time for non-urgent appointments.

The imposition of guaranteed waiting times for particular patients (e.g. cancer) would almost certainly lead to other patients suffering a longer wait. Segregation of patients into urgent slots allocated to cancer and non-cancer conditions will potentially lead to higher non-cancer waiting times.

By implication the allocation of appointment slots needs to be far more flexible than has been the case to date. As one ENT Consultant observed this implies a far more 'intelligent' booking clerk with access to Consultant diaries, knowledge of clinic arrangements and provision of forecasting tools. The ability to fill vacant slots at very short notice is an essential prerequisite arising out of randomness but has the consequence of giving some non-urgent patients access to an 'urgent' waiting time – an event usually seized upon by purchasers as evident for inequality and inefficiency!

A guaranteed 2 week cancer waiting time is almost a mathematical impossibility. This is because the weekly volumes are too small to allow reasonable prediction of next weeks demand. Indeed the high variation between Health Authorities seen within the recent statistics on cancer waits (see BMJ 15th Sep p 591) is consistent with the role of Poisson randomness (along with other factors) in the waiting time process. The role of randomness indicates that the national picture should be different in the next reporting period – although other forces are at work in tandem with randomness.

Almost certainly the unavoidable consequences of randomness will act to frustrate various government policies regarding waiting time guarantees. The correct application of such policies is to guarantee that a certain percentage of patients will be seen within the waiting time limit. The percentage not seen within the set waiting time being dictated by clinic size and resulting Poisson randomness. In this respect no single national target can be set since the performance of each clinic is a unique function of size.

The other unfortunate effect of Poisson randomness is to obscure the effect of any other forces that may be impacting upon the waiting time. The issues discussed in this paper should help clinicians and managers to work together with a clear understanding of the mathematical issues. They can then address the organisational issues upon which the mathematics has shed light.

An edited version of this article appeared as: Jones R (2001) Waiting times: quick, quick, slow. Health Service Journal 111(5778), 20-23. Please use this as the citation.

References:

1. Jones,R.P. 2001. Use of process control charts to avoid breaching routine outpatient waiting time targets. Healthcare Analysis & Forecasting, Camberley. Available from: http://www.hcaf.biz/Capacity%20Management/Microsoft%20Word%20-%20Process%20controll%20charts%20OP%20wait%20time.pdf
2. Jones,R.P. 2000. Outpatient waiting times - A pretty little sum. Health Service Journal, 111 (5740), 28 –31
3. Jones,R.P. 2000. Waiting Times - Feeling peaky. Health Service Journal, 110 (5732), 28-31
4. Beauchant,S & Jones,R.P. 1997. Socio-economic and demographic factors in patient non-attendance. British Journal of Health Care Management, 3 (10), 523 - 528.