Financial and operational risk in healthcare provision and commissioning

Dr Rod Jones

Statistical Advisor

Healthcare Analysis & Forecasting, 07890 640399 (mobile); hcaf_rod@yahoo.co.uk

Unless you understand the principle of variation in demand you will never understand healthcare resource allocation and the implied financial & operational risks.

Executive Summary

The Concept of Randomness and Variation

1.1. Contrary to our subconscious assumptions demand is not constant. Demand is variable which makes averages an unhelpful planning tool.

1.2. There are 2 sources of variation. These are:

a) That which is caused by the fact that healthcare demand operates within a complex system of short and long term cycles which means that the average is changing over time (special cause variation).

b) However, even if we knew the true average over time our actual ability to measure it – and deliver services to that level - is obscured by the fact that there is statistical variation around that average (common cause variation). This statistical based variation is described by Poisson Statistics. Hence the standard deviation (a measure of variation) is equal to the square root of the expected average

1.3. In practice the expected average is no longer accurately known because it is obscured by this statistical randomness.

What Can Be Done?

1.4. The implications for planners and operational managers is that we have to start to apply ranges with upper and lower limits rather than pretending that we know the true and precise value of demand.

1.5. Healthcare demand is actually quite small when examined on a daily or weekly basis and when split down to clinician level within an individual speciality or service. Demand on a daily, weekly or monthly basis is therefore so uncertain that the average loses its meaning for resource allocation and staff on the ground. This means that the variation is very high in percentage terms for specialities or services – and it is this phenomenon which creates a sense of lack of control.

1.6. The way to start to solve this is by the introduction of upper and lower control limits coupled with industry type control charts to complement the management process. In addition there is a need to develop approaches which enables the more flexible use of resources across teams, wards, etc.

1.7. Attempts to service variable demand using services based around an average will lead to the formation of queues as witnessed in A&E departments, and UK outpatient and inpatient waiting lists. This is only made worse when the real capacity is lower than the presenting demand. This means that optimum efficiency is actually achieved with slight over capacity (but with the assumption that staffing levels will be flexed to minimise revenue costs)

1.8. Queuing theory (which is based on Poisson Statistics) and simulation can be used to help understand the resource allocation issues.

Financial Risk in Health Care Provision & Comissioning

1.9. To the purchaser the financial risk can be contained by setting a single value contract to cover provision of health care to the population, i.e. all the risk is moved to the provider of services or by contracting at cost per case and attempting to regulate demand, i.e. risk is offset by reducing demand

1.10. To the health care provider the financial risk is a complex mixture of income (fixed or variable) and costs (with associated random variation in both fixed and variable costs!)

Introduction

Most of us grapple with this vague feeling that healthcare management is supposed to be simpler than it is. After all it seems so logical to believe that if we do extra work the waiting time will automatically reduce – but the fact is that sometimes it does not. Or we size a new unit based on forecast average workload and on some days it does not seem to be big enough. Why do these things happen?

To answer this question we need to ask another, namely, how constant is demand? The answer to this question will explain many of your unresolved healthcare resource allocation dilemmas. Put simply, demand is highly variable and part of the problem has been that you and much of the health care industry have expected it to be far more constant than it ever could be.

This leads us to a consideration of the factors which cause demand to vary and of the extent to which demand varies in practice.

Why Demand Varies?

Most of us interpret our surroundings in terms of averages. So we expect 10 GP referrals per week or 5 emergency admissions per day, etc. But what is it that determines the average and how will we ever know when the average has changed?

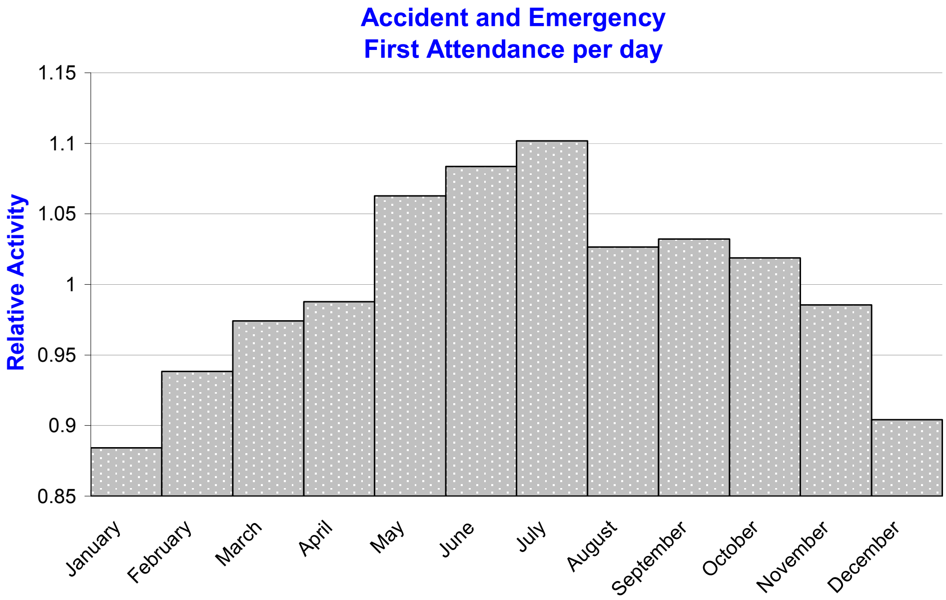
The average is not constant (special cause variation)

1. Circadian Cycles – all biological systems show circadian (i.e. 24 hour cycles), hence, for particular conditions the true incidence rate (admissions per hour) will vary with the time of day.

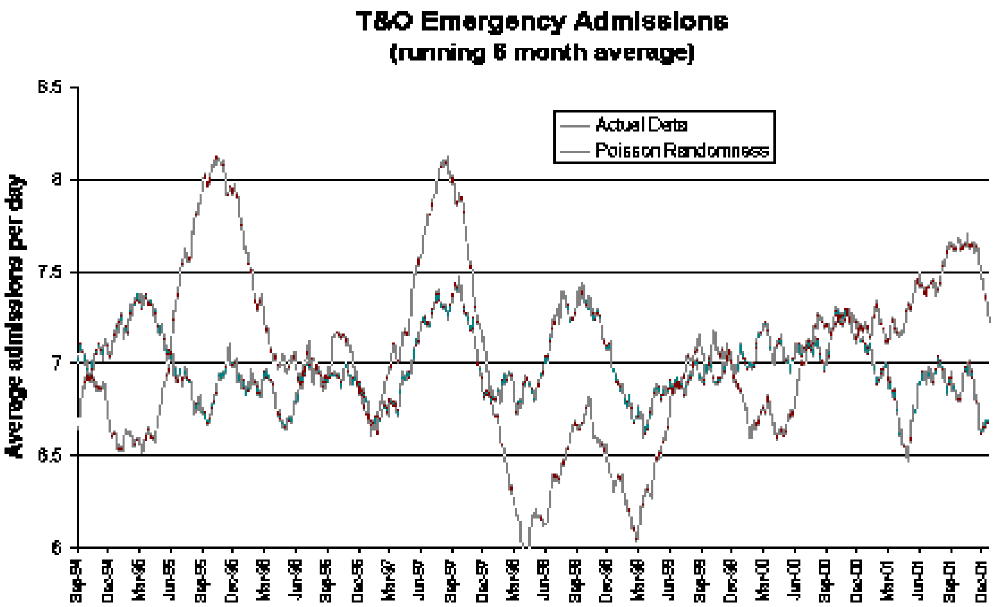
2. Daily Working Patterns - this is further complicated by GP working hours and the availability of any other supporting services. The overall effect is a distinct daily cycle in emergency admissions (greatest during working hours) competing with a working hours pattern for elective admissions.

3. Weekly Working Patterns – GP's, Social services, Home Care, etc almost all 'work' a five day week less any public holidays and hence GP referrals (including emergency admissions) likewise show distinct working day patterns. For some specialties emergency admissions peak on a Friday (prior to the weekend) and a Monday (after the weekend).

4. Seasonal Cycles – these are the annual cycles which depend on the particular condition and their interaction with the environment such as weather patterns (temperature, pressure, humidity, light intensity, pollutant levels, etc), viruses and other infections prevalent at different times of the year. Also included are the impact of school holidays and consequent flow of large numbers of people to holiday locations. See example for A&E attendance at hospital in a non-holiday season location.



5. Longer Term Cycles – the incidence of particular conditions also appears to follow longer term cycles. These longer-term trends are poorly understood. An example is attached for a large Trauma & Orthopaedic department where there is a long-term average of 7 emergency arrivals per day. The chart is interpreted by realising that each point represents the average over the previous 6 months. Hence the peak of 8.1 arrivals per day in September 1995 is an average of arrivals per day over the period April to September, i.e. roughly the spring and summer months. However, in particular years, e.g. 1998 the arrivals during this 6 month period only averaged 6.8 per day. Autumn/winter arrivals are just as variable and can range from an average of 6 to 7.3 per day. Note that the highest number of emergency admissions for a single day was 29 on 30th December, 1995 when melting snow turned to ice, i.e. over 4 times the annual average!



This chart also emphasizes the importance of taking the longer term view since if we were to base future forecasts on data from 1998 onward we may be tempted to say there was a trend upward. It is disappointing to note that most NHS organisations rarely have data going back longer that four to five years, i.e. the old style bed planning methodology and indeed the process of contracting with purchasers gave little emphasis to looking at historical trends and hence the data was not valued enough to consider keeping!

Attempts to plan based on 2 to 3 years of healthcare data is an irresponsible commitment of funds and resources

The above chart also shows the behaviour arising from simple Poisson randomness in daily arrivals. As can be seen this can lead to an apparent range in a six-month average of arrivals from 6.5 to 7.5 per day. It is interesting to note that the apparent co-incidence of some of the peaks is purely an artifact of randomness. It is a curious fact that the outcome of random events usually leads to clusters.

6. Population Demography – growth within different age bands of the population will lead to subtle shifts in healthcare demand.

7. Socological & Technological trends – these influence GP referral thresholds and the range of interventions available. For some services such as the breast clinic the volume of referral will be influenced by media coverage and even events within the latest 'soap' TV programmes.

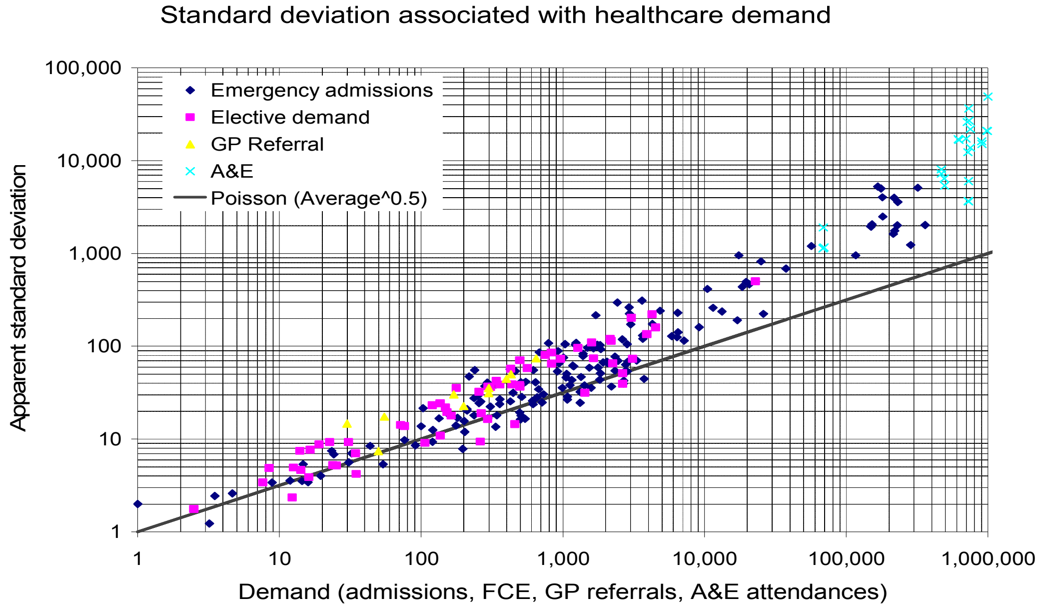
Given that there are at least 7 broad mechanisms for change in the 'average' you will now understand the need to articulate the exact specification relating to the particular hourly, daily, weekly, monthly or annual average to which you are referring.

You will also immediately appreciate the need for long-term data collection in order to determine the relative effect of the various cycles and trends.

Variation around a 'constant' average (common cause variation)

Even if we define our expected average for a given point in time there will still be variation around this average due to statistical randomness. For most healthcare demand this type of variation is described by Poisson randomness (see next figure[1]). Poisson statistics describes arrival events such as telephone calls per hour at a switchboard, customers per hour at a shop, GP referrals per week, emergency admissions per day, etc. The outcome can only be an integer value (i.e. we had 10 GP referrals last week) although the expected average can be a decimal value (i.e. our average is 8.4 GP referrals per week).

[1] The reason that real life data displays higher variation than simple Poisson randomness is related to two factors, namely; the 7 factors which influence the average and to the additional variation arising in some types of healthcare where one person makes multiple attendances each year, i.e. self harm, chronic conditions, etc. In real life it is often very difficult to determine which data in thre time series actually relates to a particular average. The reason that the data in the chart deviates to a greater extent as the numbers get larger is to do with the aggregation of dissimilar sources, e.g. trend in A&E attendance across a whole region rather than at a single site.



One highly interesting feature of Poisson statistics is that the standard deviation[2] around the average is always equal to the square root of the average. Unlike the Normal Distribution where the spread of events around the average is symmetric that of a Poisson distribution is skewed. Hence there is a tendency for more events with a value less than the average but with a tail of infrequent events at much higher than the average. This tail causes havoc to healthcare services and budgets.

This has been summarised in Table One for a range of arrival rates typical to healthcare demand. These arrival rates are typically small (at least in statistical terms) which is in contradiction to our perception that healthcare is all about large numbers of patients. Specific examples will be given to prove this, however, the point to note is that due to its size healthcare is by nature intrinsically variable and hence uncertain.

Several important points emerge from a consideration of Table One, namely:

- Even at an expected arrival rate of 20 per period it is possible (although with low probability) to get one period in which there are no arrivals.
- There are a higher proportion of periods when there are less arrivals than the expected average – don't be fooled by a series of low arrivals!
- The average arrival rate does not occur with high likelihood – the average may be the most frequent of all possible occurrences but do not therefore expect it to occur very often!

[2] The standard deviation is a measure of the variation around the average. The maximum variation is usually 3-times the standard deviation. This does not strictly apply to Poisson statistics but it is a good first approximation.

Table One: Likelihood of different outcomes given an expected average volume.

Average arrivals per period	% of periods when there are zero arrivals	% of periods with average number of arrivals	% of periods when there are fewer than average arrivals	% of periods when there are more than average arrivals
1	37%	37%	37%	26%
2	14%	28%	40%	32%
3	5%	23%	42%	35%
4	2%	20%	43%	37%
5	1%	18%	44%	38%
6	0.2%	16%	45%	39%
7	0.1%	15%	45%	40%
8	0.03%	14%	45%	41%
9	0.01%	12%	46%	42%
10	0.005%	10%	46%	44%
20	0.0000002%	9%	47%	44%

These observations lead us to a further uncomfortable question. How do we actually know the average expected arrival rate? The conventional wisdom is usually that we count the arrivals/referrals and take an average.

Managers who have studied statistics would point out that most textbooks indicate that it takes around 30 measurements to establish an accurate average. This would imply that if we measure the arrivals for 30 weeks/periods 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/period sample subject to considerable bias. However, Poisson statistics does have particular requirements and Table Two shows the accuracy obtained from one, two and three year sample periods.

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

True average per week/period	Maximum uncertainty in the calculated average given different sample sizes		
	52 weeks/periods	104 weeks/periods	156 weeks/periods
1	0.5 – 1.5	0.7 – 1.3	0.8 – 1.2
10	8.1 – 11.7	8.9 – 10.9	9.2 – 10.8
20	18.1 – 21.9	18.7 – 21.5	19.0 –21.1

This table clearly shows that the accuracy of any attempt to estimate the average declines rapidly for arrival rates below 20 per week/period, i.e. for all healthcare processes there will be high uncertainty regarding the average arrival rate. For example, at an average of 10 GP referrals per week there is a 19% uncertainty band in the calculated average using 52 weeks of data.

This has major implications to the design of 'bread and butter' computer systems within NpfIT which are supposedly suited to the needs of NHS organisations. All such systems are designed around the fallacy that demand is predictable!

If we cannot even measure the average with accuracy how then can we allocate the correct level of resources?

How accurately can we determine the underlying growth?

Most healthcare planning will involve some estimate of the underlying growth, however, simple Poisson randomness in demand is alone sufficient to confound our attempts to determine the true growth rate.

The following table gives the variation in the measured growth rate which would arise from an analysis of 5 years data assuming that the true growth rate is 10% p.a.

Table Three: Effect of volume on measured growth rate – the true growth rate is 10% p.a.

Annual Volume	Maximum	Minimum
1,000,000	10.2%	9.8%
100,000	10.5%	9.5%
10,000	11.6%	8.4%
1,000	15.2%	4.8%
100	26.4%	-6.4%
10	61.8%	-41.8%

This table highlights the relative ability of different people to 'see' the true growth rate. If we assume that at a national level the aggregated total volume is 1,000,000 per annum. At the national level, i.e. the DH, the maximum error in the estimate of the true growth rate will be 10% ± 0.2%. Imagine there are 10 strategic health authorities and so at regional level we can discern growth within the maximum bands of 10% ± 0.5%. Now imagine that within each region there are 10 PCTs and so at this level growth can only be discerned as 10% ± 1.6%. However, at any lower level than this (including almost all individual hospitals, local authroities and PBC groups) the ability to discern the true growth rate from the data is rapidly lost such that below a local volume of 1,000 per annum the estimate of growth rate is severely compromised.

Having laid the 'theoretical' framework behind demand and its sources of variation it is appropriate to apply this to a variety of financial and operational case studies. In this instance urgent GP referral, cost of a procedure and the financial stability have been chosen to illustrate the operational difficulties arising from randomness. The appropriateness or otherwise of various national targets and initiatives can then be evaluated.

At an annual volume of less than 1,000 the Poisson random variation is sufficient to dominate operational performance to such an extent that something as dramatic as 10% growth is (almost) irrelevant!

Case Study One: Urgent Referral for First Outpatient Appointment

Recent focus on achieving waiting time targets for outpatients and inpatients has led to increased awareness of the role that allocating urgent slots (for cancer, etc) plays in influencing the waiting time of both the urgent and non-urgent 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. Such apparent simplicity has probably led some to believe that 'if only those hospital managers would get their act together the waiting time targets would be delivered with ease'.

The aim of this section is to show that the randomness associated with small numbers actually makes the task almost impossible and most often leads to inefficient allocation of scarce resources. Potential solutions to this dilemma will be discussed.

The promised maximum two week wait for cancer referral has also led a perceived need for some of the urgent outpatient slots in particular specialties to be reserved for cancer patients. In this instance we now have the potential for a very delicate balancing act to ensure that all classes of patient achieve the appropriate waiting time.

Most managers would agree that there must be a better way than trial and error to achieve these simultaneous objectives. Particularly since guaranteed waiting time targets leave no room for the consequences of failure.

The apparent simplicity of allocating 2 slots per week to an expected 2 arrivals per week is shattered by the fact that Poisson statistics tells us that outcomes other than the average are highly likely. In order to make the following discussion of practical relevance we must first establish the level of urgent referrals received by consultants.

The situation regarding the general level of urgent appointments is illustrated by reference to the Royal Berkshire Hospital (around the 20th largest Trust in the UK). The two clinics having the highest provision of urgent slots are Cardiology where at the time there were 38 urgent slots per week jointly managed by two Consultants, i.e. around 19 per week per consultant. These slots were rarely filled and closer to the clinic date they were re-allocated to non-urgent patients. This results in some routine patients receiving a very short wait since they are placed into an 'urgent' slot while other routine patients wait a longer period of time in one of the standard routine appointment slots. This is a direct consequence of the need to be operationally relevant, efficient and not wasting scarce resources!

The next highest provision is for a General Surgeon specialising in GI surgery where there are 9 urgent slots per week. In contrast the average across all specialties is only 2 per week and the most common value is 1 per week.

For cancer referrals the largest weekly average is for the combined upper and lower GI tract where a large acute hospital would receive around 10 to 20 per week. This is spread over a number of consultants in both General Surgery and Gastroenterology and hence the average per consultant is usually less than 5 per week. In most instances the highest average of new cancer referrals per consultant is less than 2 per week.

Having established the boundaries we can use Table One to investigate the effect of Poisson randomness on such small number events. How will Poisson randomness influence attempts to efficiently allocate scarce resources?

Table Four: Hypothetical clinic where number waiting at start of year is sufficient to avoid wasted clinic slots due to lower than average referrals received during the year.

Average referral rate		**Referrals actually received in year**		**Number Waiting**	**Waiting time (weeks)**		
Per Year	Per week	Maximum	Minimum	Start of year	Start of year	Last day of year (Maximum)	Last day of year (Minimum)
1040	20	1105	975	65	3	7	0
936	18	997	875	61	3	7	0
832	16	890	774	58	4	7	0
728	14	782	674	54	4	8	0
624	12	674	574	50	4	8	0
520	10	566	474	46	5	9	0
416	8	457	375	41	5	10	0
312	6	348	276	36	6	12	0
260	5	293	227	33	7	13	0
208	4	237	179	29	7	15	0
156	3	181	131	25	8	17	0
104	2	124	84	20	10	20	0
52	1	66	38	14	14	28	0

In this respect most consultants or managers would not wish to have empty clinic slots since this is clearly a waste of resource. To avoid this possibility we could theoretically set up a clinic with sufficient patients waiting at the start of the year to avoid the possibility of lower than average referrals leading to empty clinic slots toward the end of the year. However at the same time we may actually get more referrals than the average and hence the waiting time could increase rather than decrease. Table Four explores this dilemma for various levels of urgent referral where the maximum and minimum referrals are at the 95% confidence intervals, i.e. higher and lower numbers of referrals will only occur on 5% of occasions. At the top of the table we have the equivalent to the Cardiology clinic while at the bottom are the most common incidence of 1 per week.

It would appear that Poisson variation defeats all 'steady state' attempts to efficiently allocate scarce resources within the context of attempting to deliver a clinically acceptable waiting time.

Due to randomness in the arrival of urgent referrals we see that the minimum possible urgent wait to avoid wasting scarce resources is three weeks (assuming 20 referrals per week) but that this could lead to a maximum wait of seven weeks due to higher than average arrival of referrals. **Indeed it can be seen that all possible options in Table Four are incompatible with a four week maximum outpatient wait.**

New methods of management are required

It would seem that our inability to know the true average and the confounding effect of random variation around this (estimated) average has led us into a dead end. It has however explained why it is impossible to allocate the correct number of urgent appointment slots. How do we extricate ourselves from this dilemma?

The steps to avoiding this dilemma rely upon our realising that we need to respond to randomness as it occurs rather than attempting to allocate the 'correct' number of clinic slots ahead of time.

There are three approaches which can be employed. In the first we allocate the maximum possible number of first or urgent slots that could arrive in the given period (Table One). On most occasions this leaves us with many empty appointment slots which can be filled with newly arriving routine appointments or left free for other consultant duties such as administration, research, etc. This approach assumes a high degree of flexibility in the booking of patients and also assumes moderate to high overcapacity in terms of clinical & supporting resources. These conditions are rarely met in practice; however, if at all possible, it does lead to very low waiting times.

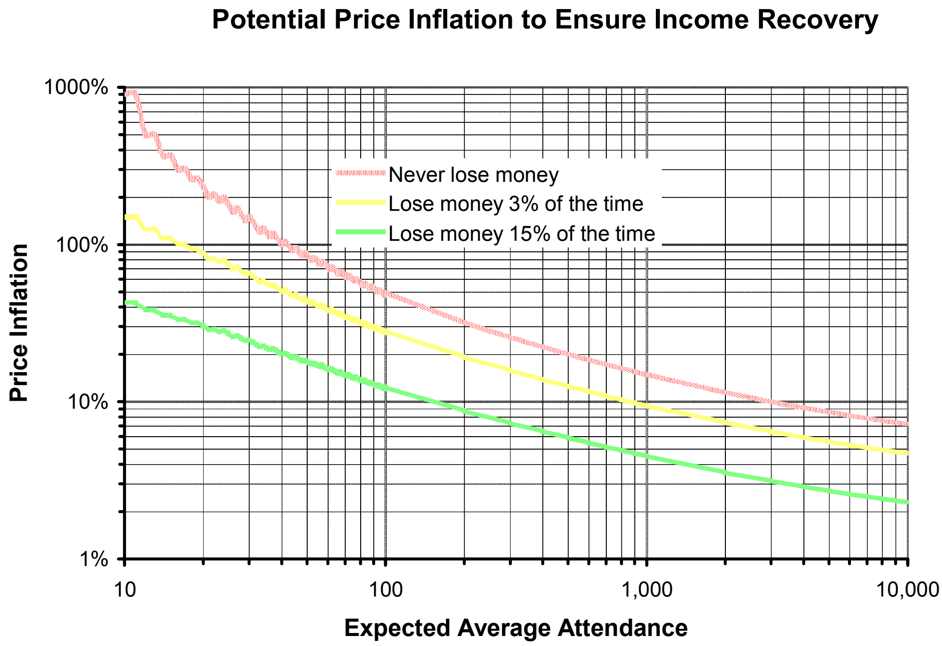
In the second appraoch we allocate the weekly average plus, say, one half to one standard devaition (i.e. the standard deviation associated with the annual total divided by 52 weeks) as the number of available slots. Under this scenario additional slots are only required on rare occasions and these can be incorporated as over-booking. Less excess resource is needed and low waiting times are maintained by virtue of excess capacity over demand.

The third approach is more pragmatic and uses process control charts to help us decide when to provide extra capacity to bring the waiting time back 'in control'. This minimises total cost and more closely approximates current working arrangements.

Case Study Two: What price to charge for a procedure?

If demand is uncertain then what price needs to be charged to cover both fixed and variable costs?

The unfortunate answer is much higher than anyone in the NHS is allowed to charge! This partly explains why non NHS organisations such as Kaiser Permanente or private sector hospitals can 'do a better job' of managing healthcare finances, i.e. they are not bound by UK rules of finance for publically funded organisations which are in direct contradiction to the demands made by variation in healthcare demand.



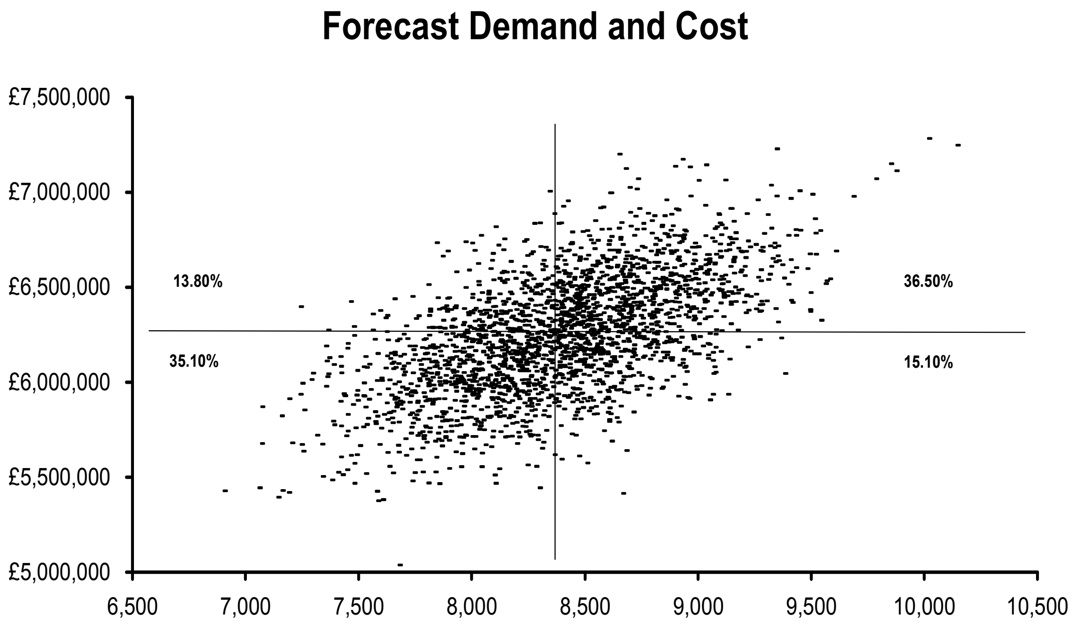
This also suggests that the move to PbR may have some unexpected adverse side effects based on the size of the healthcare provider, i.e. smaller organisations will tend to be less financially stable.

Case Study Three: Income received under PbR funding?

The following chart details the outcome of a computer simulation of the income received (and hence the cost borne by a PCT) from delivering a range of acute outpatient clinics. The chart graphically shows the huge variation seen in 2,000 possible outcomes <u>for a single year</u>. In this simulation the demand for each clinic is allowed to vary according to Poisson randomness while the price per patient is set at the outpatient specialty tariff. On the x-axis is all possible outcomes for total volume (across all specialties) in a single year while on the y-axis is the income associated with these total volumes. This simulation merely shows the expected variation in income while the cost associated with delivering theses services can be modelled by a slightly more sophisticated simulation scenario where fixed and variable costs are separated out.

What the simulation does show is that high volume-high income outcomes are the most likely (but only account for 36% of all outcomes) and that high volume-low income or low volume-high income outcomes are less likely. The simulation also demonstrates the almost impossible task faced by almost all finance directors – attempting to manage randomness in income as well as volume sensitive variation in costs!

Once again new management strategies are required and the area of financial management is almost begging for a new approach utilizing industrial style process control charts to track variation in demand, so-called fixed costs, semi-variable variable and step costs.



Case Study Four: What tolerance is required on a contract

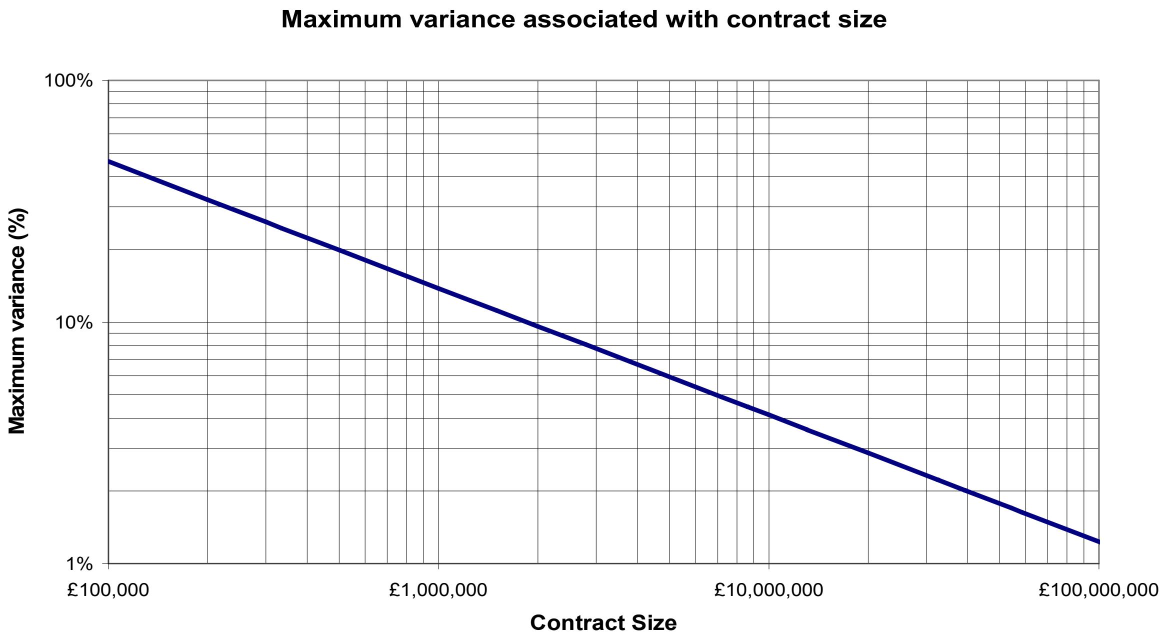
Consider a contract to supply healthcare services to a defined population such as a PCT. Such a contract should contain a tolerance to account for over performance or to trigger re-negotiation.

An almost intuitive approach would indicate that the % tolerance for a larger contract should be smaller than the % tolerance on a small contract. Monte Carlo simulation can be used to determine the approximate relationship between volume and % tolerance.

The attached chart is the output from one such simulation. The tolerance presented in the chart is felt to be conservative for several reasons:

1. Poisson was assumed to be the only cause of variation – in practice healthcare variation is usually 2 to 3 times higher than simple Poisson randomness.

2. The simulation used specialty averages – had the simulation been taken to the level of HRG the variation would have been higher.
3. The simulation assumed a balanced portfolio of general acute specialties – tertiary services are likely to be more variable.



However, in spite of the reservations that the above tolerance may be an underestimate we can see that in relative terms a 20% tolerance would apply to contracts of around £500K, a10% tolerance would apply to contracts of value £2M, a 5% tolerance will apply at £7M, 4% at £10M, 3% at £20M, 2% at £40M and 1% for contracts in excess of £100M.

Since the inherent variation is high it is not recommended that contracts be monitored at specialty level or using activity. Total contract value is the preferred point to exert overall control.

Conclusions:

The NHS has suffered greatly from a simplistic view of demand. This is most aptly illustrated in attempts to plan 'activity' in the LDP. Organisations are asked to submit the 'number' and cost of next years activity. Indeed the assumptions behind NHS contracting demonstrate a similar view, namely, it must be easy to forecast demand therefore it must be easy to put fixed numbers into a contract or a spreadsheet!

Hopefully this paper has whetted your appetite to think in a different way. How will this new thinking influence the way you manage capacity, employ staff, acquire new assets and run your service?

Healthcare Analysis & Forecasting uses a range of forecasting tools, simulation software and appropraite models to solve resource allocation and financial risk issues within healthcare, namely, how many beds does a specialty or hospital need, how do we allocate resources to guarantee waiting time targets, how much activity (or more correctly what upper and lower limits) needs to be in a contract to guarantee achieving a target, what are the seasonal profiles behind activity and waiting lists, how stable is our financial position, how big should our contingency budget be, etc.

Additional analytical tools include geo-demographic analysis to locate the feasible or best location for new healthcare services, potential changes in patient flows in a choice environment, or the changes in patient flows should service provision be changed (i.e. relocation to a new site or a new competitor enters your locality) and a tool which re-interprets the resource allocation formula in a more meaningful way at specialty and HRG level to answer questions such as, is our spend on Orthopaedics too high or are our GPs making too high a volume of referrals relative to our population and its demographic, ethnic and socio-economic profile?

Contact Details:

Dr Rod Jones
Statistical Advisor
hcaf_rod@yahoo.co.uk (email)
07890 640399 (mobile)