Gooroo for NHS

1

Non-elective bed occupancy **Introduction**

JANUARY 2019

What should your non-elective bed occupancy be? It's an important question because it is a huge driver of staffing, cost, emergency department waiting times, elective cancellations, and patient safety.

Yet much of the NHS uses 'magic numbers' for bed occupancy. The most popular is 85 per cent, which relates to a specific spreadsheet simulation but is widely cited as something for every hospital to aspire to. Then if there aren't enough beds, the hospital tries to operate at higher bed occupancy without appreciating the risk they are accepting.

This paper explains why the 'right' bed occupancy is different for every pool of beds, why getting it right matters, what needs to be taken into account to determine it, and why we should be talking about risk instead.

The second paper will explain how bed occupancy can be calculated, the third will examine some management approaches to reducing bed occupancy and risk, and the fourth paper will look at the implications for elective beds and hospital configuration.

gooroo

What is bed occupancy?

The term 'bed occupancy' broadly means the proportion of available beds that are occupied.

'Available' means the bed is staffed and ready for use at short notice, and 'occupied' means it is being used by a patient. So far so good. And yet the exact meaning of 'bed occupancy' varies depending on the context:

1) If you take a snapshot of the current available beds, then bed occupancy might be the proportion that are currently occupied. For instance, you might take a snapshot of acute beds beds at 8am, and find that there are 200 beds available of which 180 are occupied by patients. In this snapshot, 'bed occupancy' is 90 per cent.

2) If you are calculating the long-term number of available beds required in the hospital then the usual formula is:

$$\text{available beds} = \text{average discharges per day} \times \text{average length of stay} \div \text{bed occupancy}$$

...but even this simple-looking formula can have different meanings depending on how 'average length of stay' (ALOS) is calculated:

a) If the length of stay (LOS) of any patient is calculated simply as the difference between their admission date and their discharge date, then LOS is the number of midnights that the patient was in a bed. It follows that ALOS is the average number of midnights, and therefore 'bed occupancy' is the average proportion of beds occupied at midnight.

b) If however LOS is the exact number of days, hours and minutes between each patient's admission and discharge, then 'bed occupancy' is the average proportion of beds that are occupied throughout the period. This is the definition we will be working with in these papers.

3) If you know how many beds you are supposed to have, then you can calculate 'bed occupancy' by rearranging the above formula to:

$$\text{bed occupancy} = \text{average discharges per day} \times \text{average length of stay} \div \text{available beds}$$

Although the arithmetic is the same, the meanings can vary depending how 'available beds' is defined (and here the language can become rather loose). For instance:

a) If you have funding for 100 medical non-elective beds, and they are all occupied with a further 20 patients spilling out into other parts of the hospital, then you might say that your bed occupancy is currently 120 per cent.

b) Similarly, if you have funding for 100 beds and on average 100 beds are occupied, then you might say that your bed occupancy is 100 per cent. But admissions and discharges vary, so sometimes you will have spare beds and at other times patients will spill out elsewhere in the hospital (or be turned away). Usually the probabilities will be symmetrical, so 100 per cent bed occupancy implies that the risk of running out of beds is about 50 per cent.

This brings us to the question of risk. What is the acceptable risk of running out of beds in each bed pool? And how does risk vary with bed occupancy?

The purpose of a low enough bed occupancy

All the above definitions of bed occupancy have their uses when it comes to measuring bed occupancy, but none of them shed much light on what bed occupancy *should* be (although in the final example it is clearly too high). That is because the above formulae are based on averages, whereas the 'right bed occupancy' is a risk-based concept that depends on variation.

Imagine for a moment a fictitious elective service, where every patient is booked to arrive on the ward at precisely the moment the bed becomes available following the discharge of the previous patient. This hypothetical ward can run at a bed occupancy of nearly 100 per cent, because everything is completely predictable and managed. A more realistic assumption for real wards might be that elective beds could run at a bed occupancy of 100% minus the Did Not Attend rate.

But **non**-elective beds are not completely predictable. There are certainly some patterns – respiratory admissions tend to be higher during winter, and medical admissions lower at weekends – but overlaid on top of those typical variations is a large amount of sheer randomness.

Even if we have plenty of beds, there is some probability that a surge in admissions will fill them all, and then we will be faced with the familiar, unpalatable options of holding new arrivals in the emergency department, allowing them to spill out into other areas of the hospital, cancelling elective patients, or turning patients away by diverting the ambulances. In other words, there is some probability of running out of beds.

So if we want to determine the right bed occupancy for our beds, we need to distinguish between the variation that can be predicted and managed ahead for, and the variation that cannot. We can plan and manage the former by adjusting the numbers of available beds, and by scheduling elective patients around the predictable peaks. And then we need bed occupancy to be low enough to absorb the latter.

Forecasting range

The NHS works to predict and plan its capacity requirements over various timescales, such as:

- many years ahead for the strategic planning of capital and consultant medical staff,
- a year or two ahead for longer-range financial planning,
- months ahead for some non-permanent medical staff and their associated sessional commitments,
- about six weeks ahead for annual leave and associated locum cover,
- up to few weeks ahead for extra sessions, and
- days ahead for bank and agency staff on the wards and for short-notice locum cover.

In general, short-range forecasting is more accurate than long-range forecasting, as we know from our experience of weather forecasts. But longer-range forecasts are still needed because buildings and staff are long-term investments that take time to change. So in practice the NHS takes a blended approach to planning, in which the

broad shape of services is planned and managed for the years and months ahead, and then capacity is operationally fine-tuned for the coming days and weeks.

The principle we will follow in these papers is that bed occupancy only needs to be low enough to absorb the genuinely unpredictable variations in demand, i.e. the variation that remains after planning and management has been done over all the timescales listed above.

A possible objection to this principle is that longer-range planning has been ineffective in recent years, because the NHS has not had enough resource to act on those plans by implementing the necessary capital and staffing. However it does not follow that any calculation of bed occupancy needs to allow for this long-term planning risk. That would lead us to conclude that bed occupancy should have been set lower, many years ago, in anticipation of today's capacity crises – a proposition that is both circular and self-defeating. By the same logic, bed occupancy is not intended to apply to the annual planning process either.

The next question is: how 'short' is short-range planning? The answer depends on the practical response to those plans; if today is turning out to be busy then it may be already be too late to do anything about it, but if next week is predicted to be busy then perhaps something can be done.

The management options available to adjust capacity at the shortest notice and at acceptable cost include:

1) increasing or decreasing the amount of elective care that shares the same bed pool; and

2) increasing or decreasing the number of available beds, e.g. by adjusting bank and agency staffing levels.

Looking at the first option, we know that elective patients should be given 'reasonable notice' of surgery, which is usually defined for routine patients as 3 weeks' notice in England (or 48 hours for cancer) and 1 week in Scotland; although shorter dates can be considered 'reasonable' if the patient accepts. So we need to be able to forecast that far ahead to schedule extra elective care effectively around the non-elective peaks. (Although elective care can be cancelled at much shorter notice.)

Looking at the second option, bank and agency staffing levels can be adjusted at short notice. But if the hospital does not usually have a significant number of such staff then there will be limited scope to reduce available bed capacity this way (although there may be scope to increase it).

There may also be a minimum number of beds that can be opened or closed at once, depending on how the hospital is managed. If whole wards are either 'open' or 'closed', then changing the number of staffed beds will only be possible for large (e.g. seasonal) changes in bed numbers. On the other hand, if adjustments can be done at bay or single-bed level then more fine tuning will be possible.

Depending on how beds are managed, then, we could regard non-elective demand as being 'predictable' if it can be forecast 1 to 3 weeks ahead, and if we always have an up-to-date forecast upon which we can act. If forecasts are updated less frequently, or if we cannot adapt elective scheduling and/or staffing at short enough notice, then we may need to forecast further ahead; however then our forecasting will be less accurate, so more of the variation in demand will be unpredictable, and either bed occupancy will need to be lower or the risk of running out of beds higher.

Bed pools

If you toss 10 coins, the odds that 70 per cent or more will be 'heads' are better than 1 in 6.

If you toss 100 coins, the odds that 70 per cent or more will be 'heads' are worse than 1 in 25,000.

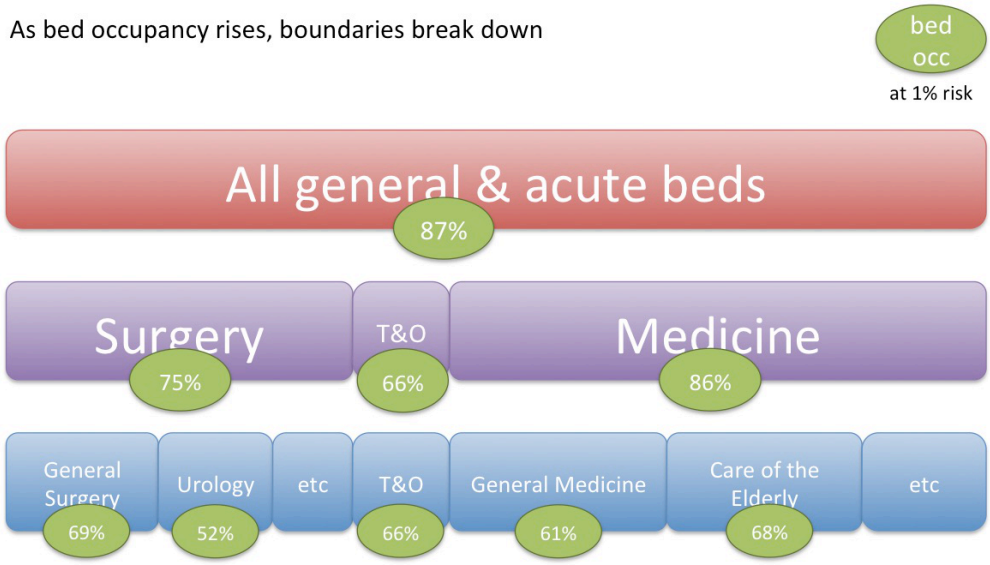
Similarly, large bed pools can absorb unpredictable variations in non-elective demand more easily than small bed pools – fluctuations are more likely to average out in a large bed pool.

What counts as a 'bed pool' in a real NHS hospital?

Acute hospital beds are often allocated in some detail to individual specialties (e.g. respiratory, or geriatric medicine), as part of the financial planning process. However if each individual specialty had to absorb all of its non-elective admissions within its own beds, then it would need to operate at quite a low bed occupancy. This is shown by the lowest layer in the figure below (the example bed occupancies shown were calculated for a 1 per cent risk of running out of beds in a particular general acute hospital).

In reality, most hospitals do not have enough beds for that. So several specialties end up sharing a larger pool of beds between them. This combined bed pool is better able to absorb variation and can run at higher bed occupancy, which means that fewer beds are needed overall. A typical acute hospital might have an bed pool that extends across most of medicine, with separate bed pools for surgery, orthopaedics, paediatrics, gynaecology, maternity, and critical care. Within each bed pool, the separate specialties keep to themselves as much as they can, but frequently spill out into other beds within the pool in a reasonably controlled way. This is shown in the middle layer below, for surgical and medical beds.

But if bed occupancy runs very high, then even those boundaries start collapsing and the hospital is effectively turned into one enormous bed pool. Medicine starts spilling over into surgical beds and vice versa, female surgery into gynaecology, and surgery into orthopaedics, and patients spill out into other areas such as day surgery and corridors. This creates significant clinical risks for non-elective patients who end up in the wrong place, it causes cancellations and delays for elective patients, and is a scenario that everybody wants to avoid. In the top layer of the example below, bed occupancy has not changed very much for medicine, but it has changed for surgery and this is where the extra beds came from as the pressures increased.



The risk of running out of beds

The bed occupancy figures in the example above were based on a 1% risk of running out of beds. But the level of acceptable risk will vary from one bed pool to another, because the consequences of running out of beds are different.

For instance, if the main adult ward block of a general hospital runs out of beds, then the consequence might be cancelling some routine elective surgery; this is certainly an inconvenience, a poor service to patients, and a waste of costly resources, and it may increase clinical risk but is not usually an immediate threat to life. However if the maternity unit were to run out of beds, then women in labour might be turned away; a serious clinical risk.

It would therefore be sensible to plan maternity bed occupancy for a much lower risk of running out of beds (say, 0.1% of the time) than would be acceptable in the main ward block (which might plan to run out of beds several percent of the time). These levels of risk can be illustrated by turning them into frequencies:

0.1% is about 9 hours per year
0.3% is about 2 hours per month
1% is about 2 hours per week
3% is about 5 hours per week
10% is about 2 hours per day
30% is about 7 hours per day

So rather than talking about bed occupancy, it would be better to talk about the acceptable risk of running out of beds. When this risk has been agreed, it is then a technical exercise to convert that risk into bed occupancy by analysing the variation in non-elective demand (as the next paper in this series shows).

But if bed occupancy is allowed to rise without a good understanding of the relationship with risk, then the hospital will be prone to operating at unexpectedly high levels of risk without realising it. This is the position at many acute hospitals across the NHS today.

Conclusion

'Bed occupancy' broadly means the proportion of available beds that are occupied. However the exact meaning depends on the timescales being considered and what is meant by 'available beds'. In these papers we will use 'bed occupancy' to mean the average proportion of beds that are occupied continuously over time.

Larger bed pools are better at absorbing variation than small ones. When bed occupancy is high, the boundaries between bed pools break down to form larger bed pools, so patients sometimes end up in the wrong place which creates clinical risk.

Discussion about bed occupancy should be replaced with discussion about the acceptable risk of running out of beds. Once the acceptable risk has been agreed for each bed pool, it is then a technical exercise to calculate the corresponding non-elective bed occupancy.

Some of the variation in non-elective demand is predictable, and can be planned and managed. Then bed occupancy needs to be low enough to absorb any other variation that cannot be planned and managed, at the acceptable risk of running out of beds. It follows that better planning and management allows a bed pool to operate at higher bed occupancy (and therefore with fewer beds) because less of the variation is unpredictable.

Because the variation in demand differs from bed pool to bed pool, the relationship between bed occupancy and risk must be calculated separately in each case.

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

v1.0 22 Jan 2019