Adapting Lean Processes for the Hospital/Surgical Environment

Jaideep J Pandit and Meghana Pandit

1. Consultant Anaesthetist & Senior Clinical Lecturer, John Radcliffe Hospital and University of Oxford, United Kingdom
2. Consultant Gynaecologist, Milton Keynes Hospital Foundation Trust, Milton Keynes, United Kingdom

1. Introduction

Hospital operating theatres are a focus for cost reduction, especially as expenses can run into billions of dollars (in the United Kingdom’s National Health Service, theatres have been estimated to cost >£1 billion). About 46% of patients discharged from hospital have undergone surgery (Gordon et al. 1988; Audit Commission, 2003; Berwick, 2005; Cegan, 2005). Yet, cancellation rates can reach up to 20% and waiting lists for surgery exist in many countries (Gauld & Derrett, 2000; Buhaug, 2002; Bellan, 2008).

The concepts of ‘Lean’ or ‘Six Sigma’ thinking have shown great promise in industry, because they seek to reduce variations in inputs (eg, in quality of raw materials or steps in manufacturing processes), which increases efficiency and reduces costs. Although attempts have been made to apply these concepts to healthcare, it is not proven that their introduction has made progress or reduced costs (Vest & Gamm, 2009; Pandit et al., 2010). Therefore, these ideas may need considerable adaptation for the healthcare setting. This article focuses on three approaches to help understand the problems, and therefore to solve them: first, the notion of matching surgical capacity to demand for surgery; second, the idea of what constitutes ‘efficiency’ and ‘productivity’ in a surgical list; and third, we describe how effective planning of a surgical list using quantitative data reduces over-runs and patient cancellation. Together these ideas demonstrate how ‘Lean’ is suitably adapted to the existing circumstances in the surgical-anaesthetic setting.

At the outset, it is important to distinguish between operational, strategic and tactical decision-making in relation to operating theatre management. Operational decisions concern day-to-day local problems (eg, late starts or transportation problems). The relevant solutions are hospital-specific and may not apply to all hospitals and set the environment in which the organisations function. Strategic decisions concern the global direction/delivery of the service (for example, socialised vs private healthcare, relationships between funders and providers, etc). These decisions affect all hospitals. Tactical decisions are short-to-medium term concerning service planning to implement the strategic decisions (for example, optimum models for theatre scheduling, theatre allocations etc). Tactical analyses apply to all hospitals working within the same strategic environment.

Whereas Lean/Six Sigma approaches are usually focussed upon processes within a patient’s journey in hospital (see: http://www.institute.nhs.uk/) and so are traditionally considered
to function at an operational level, we believe that there is scope for translating Lean ideas to the tactical level, in a quantitative approach to demand-capacity and list planning.

2. Choosing the right surgical capacity for prevailing demand

It is a common problem to try ascertain how many hours of operating a particular surgical team needs per week. In some countries (eg, US) theatre time is not scheduled as block-time, but variable for each team (specialty) depending upon how many referrals it receives. Thus, ‘capacity’ is not a fixed quantity, but adjusted to match a variable demand and also to create incentives. In these settings, hospitals can modify their capacity for certain surgical services as a means to compete for business (Dexter & O’Neill, 2004; Pandit & Dexter, 2009). In other countries (eg, UK), surgical capacity is largely agreed and fixed long in advance and regarded merely as a passive means to cope with an ever-present demand in a socialised system of healthcare (Pandit et al., 2010).

If surgical ‘capacity’ matches (or exceeds) ‘demand’ for surgery, then patients are promptly treated and there are no surgical waiting lists. However, both the terms ‘demand’ and ‘capacity’ require some mathematical explanation. Only then can be understood what is meant by ‘matching’ these two quantities. In the context of healthcare - and surgery in particular - ‘time’ is more relevant a measure than is the absolute number of patients. Thus, demand is best understood as the total minutes or hours required for the surgical procedures, and not simply by the number of patients booked for surgery. It is notable that this ‘demand’ arises from the patients the surgeon sees in the outpatient clinics each week. Correspondingly, ‘capacity’ is the weekly operating time available to the surgeon.

![Demand generated in a surgical clinic over 10 weeks.](image)

If demand were known and constant then the problem would be simple. However, the surgeon sees a different number of patients each week, needing different procedures.
Therefore, the variation in demand (not just the mean demand) influences any mathematical analysis. However, whenever we introduce ‘variation’ into the equations, and whenever that ‘variation’ is unpredictable, then we have to deal with concepts like ‘likelihoods’ and ‘probability’.

Fig. 2. A: Histogram for expanded set of data in Figure 1, showing the three possible capacity levels that could be set. The area under the curve to the left of each line yields the probability of meeting the demand at the chosen level of capacity. B: Integral of Figure 2A. The probability is now shown on the y-axis for the levels of capacity chosen.

Figure 1 shows a variable demand (a surgeon books a variable number of hours of surgery per week). If capacity is set too low (line 1), a waiting list will result. If, however, set too high, then demand is reliably absorbed, but there is considerable potential waste of
resources (line 2). Between these limits, capacities absorb demand in some weeks but not in others (line 3). What is the correct level of demand?

One way of approaching this problem systematically is to consider the histogram of demand generated from the clinic activity (Figure 2A). We simplify the considerations to a normal distribution, but similar calculations can be performed for non-normal distributions (Pandit et al., 2010). For any capacity hypothetically set (eg, at lines 1, 2 or 3 in Figure 2A that approximately correspond to these lines in Figure 1), we can now estimate the likelihood of absorbing demand as the area under the curve to the left of each hypothetical vertical line for capacity (Figure 2A). In other words, mathematically integrating the area under the normal curve gives us a probability density function (Figure 2B) for the relevant levels of capacity. This density function tells us the likelihood of absorbing that demand on a regular basis for a chosen level of capacity. Whether a likelihood of 70%, 80% or 90% is chosen depends on several factors, of which one is ‘waste’.

There are at least two concepts of ‘waste’ in the current context. One is the notion that money is spent with little gain. With reference to Figure 2B, increasing surgical capacity by a given quantum (eg, by 1000 min/wk) will cost a certain amount of money. Theatre costs have been estimated as between £12-20.min⁻¹, so this would costs £12,000 – £20,000.wk⁻¹ per theatre (Abbott et al, 2011). If surgical capacity is increased by this amount from 2,000 min.week⁻¹ to 3,000 min.week⁻¹, this would increase the proportion of weeks in which demand was met by ~40% (from ~40% to ~80%; Figure 2A).

However, increasing surgical capacity by the same quantum from 3,000 min.week⁻¹ to 4,000 min.week⁻¹ increases the probability of meeting demand by only ~10% (from 80% to ~90%; Figure 2A). These diminishing returns on investment represent a form of waste, and must be appreciated by all organisations that wish to be ‘lean’

---

Fig. 3. The inverse relationship between utilisation and waste (bold lines; read from left axis) for given levels of capacity for the data shown in Figures 1 and 2. Superimposed (dashed line; read from right axis) is the same plot as in Figure 2B.

The second notion of ‘waste’ is that of unused capacity. For any given level of capacity for the hypothetical demand, we can calculate the proportion of capacity that will be wasted.
The proportion of time wasted rises as capacity is increased (while the proportion of time utilised correspondingly declines) (Figure 3). A very low capacity of ~ 1,000 min.wk⁻¹ is associated with almost no wasted time and there is high utilisation. But this level of capacity will absorb demand in very few weeks, causing a huge rise in the waiting list (Figure 3). On the other hand, capacities of ≥4000 min.wk⁻¹ which absorb demand every week are associated with ≥50% of time wasted (ie, <50% time utilised; Figure 3). There is therefore an inevitable trade-off between choosing a capacity which reliably absorbs prevailing demand, and choosing a capacity that minimises waste (Macario, 2010). The optimum balance between these greatly depends on the local, social, or political priorities.

In a flourishing economy, it may be possible to absorb all demand despite a degree of waste. In times of economic hardship it may be necessary to minimise all waste and accept prolonged waiting times or waiting lists as one of the prices to pay for potential economic recovery. In other words, achieving too high a utilisation is as bad as achieving too low a utilisation (Table 1). In healthcare, the ‘lean’ option not simply attaining the highest possible utilisation; rather it is achieving a balance between utilisation of resources and several other factors, one of which is waste.

<table>
<thead>
<tr>
<th>Utilisation</th>
<th>Implication</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>Not realistic and/or implies insufficient capacity</td>
</tr>
<tr>
<td>90%</td>
<td>Exceptional; may not be sustainable</td>
</tr>
<tr>
<td>80%</td>
<td>Implies a good scheduling algorithm and is achievable</td>
</tr>
<tr>
<td>70%</td>
<td>Acceptable, but room for improvement</td>
</tr>
<tr>
<td>50%</td>
<td>Implies poor management or profligate approach (ie, theatre capacity possibly excessive)</td>
</tr>
<tr>
<td>30% or less</td>
<td>Unacceptable, implying very poor theatre management</td>
</tr>
</tbody>
</table>

Table 1. Degrees of theatre utilisation and their implications.

3. Measuring efficiency in a surgical list

Another aspect of utilisation is the proportion of time used within a single surgical list (say, each list is of 8 hours duration). Ideally, the amount of time wasted in non-productive activity (eg, waiting for the patient to arrive, opening equipment packs, etc) should be minimised and as much of the 8-hour list as possible should be spent in productive anaesthetic-surgical tasks. Yet, the list should not over-run its allotted time as this is expensive (due to overtime payments and unbudgeted consumables), and unplanned over-runs can disrupt other aspects of the clinical service, including emergency work. An over-running list can also result in patient cancellation (ie, where patients are still waiting on the ward to be called to theatre and have their surgery but the original end-time of the list is long past and staff cannot or will not stay despite offer of overtime payments). Indeed, over-runs are possibly the main cause of cancellation (Pandit & Carey, 2006; Pandit et al, 2007). Given these considerations, what is an ‘efficient’ surgical list?
Simple list ‘utilisation’ can be excluded as an appropriate measure of efficiency for several reasons. First, optimal list utilisation differs among surgical specialties. Not all specialties can achieve equally high utilisation (Dexter et al., 1999). For example, orthopaedic surgeons specialising in joint replacements are more likely to know in advance which cases they are doing and so utilise their time fully as compared with, say, cardiac surgeons who only know a day or two in advance which patients admitted with unstable angina will need revascularisation. Second, measured utilisation can be highly variable from week to week (eg, from 38% - 85%; Dexter et al., 2003) so questions arise as to whether means or ranges are most relevant. Third, utilisation is irrelevant when some specialties are fostered to promote the general status of the hospital. For example, a new service such as robotic prostatectomy will naturally have low utilisation initially but will serve the hospital well over time (Macario, 2010). Finally, utilisation figures can be artificially maximised by poor practice. For example, teams can slow down simply to occupy the theatre for the ‘target time’. Or, since it is easier to fill the list with shorter cases than longer cases, the former may be preferentially booked at the expense of the latter (Macario, 2010).

By contrast, all can agree that the following sentiment encapsulates the notion of good efficiency on a surgical operating list: a theatre is used most efficiently when as much of the time available is utilised, when there are no over-runs and no patients are cancelled (Widdison, 1995). This can be expressed mathematically as:

Equation 1:

\[
Efficiency = \left[ \left( \frac{\text{fraction of scheduled time utilised}}{\text{fraction of scheduled time over-running}} \right) \right] \times (\text{fraction of scheduled operations completed})
\]

If an 8 hour list finishes in 6 hours, then the ‘fraction of scheduled time utilised’ is 0.75 and the ‘fraction of scheduled time over-running’ for this same list is zero. If an 8 hour list over-runs by 2 hours, then the ‘fraction of scheduled time over-running’ 0.25, and the fraction of scheduled time utilised for this list = 1. Thus the first two terms of the equation operate in a mutually exclusive manner: i.e. a single list cannot be both under- or over-utilised at the same time. If 4 of 5 patients scheduled on a list are completed and one is cancelled, the ‘fraction of scheduled operations completed’ is 0.80. The formula therefore theoretically yields a result for efficiency ranging from 0 to 1.0 (or 0–100% if this result is multiplied by 100). The value of 100% is obtained when all booked cases are complete at the scheduled time, which is our sense of perfect efficiency. The formula can also give ‘credit’ for a list that completes its own booked cases early (e.g. four cases) and accepts and completes extra cases (e.g. a fifth case from another list). Thus a number > 1 in the last term (i.e. the fraction of patients completed = 1.25 for this example) could translate as an efficiency > 100% for that particular list.

The formula can be shown graphically where cancelled operations ‘set the envelope’ for the maximum efficiency (i.e. there are ‘isopleths’ set by the cancellation rate) and efficiency increases to a maximum if the list finishes at the scheduled list end-time, but then declines thereafter (Figure 4). Plotting a list on this graph will show whether inefficiency results from under-utilisation (point A) or over-running (point B) or cancellation (point C). Combinations can also be readily visualised (point D is an under-running list with a cancellation; point E is an over-running list with a cancellation). This measure of efficiency is being recognised as a useful standard, with efficiencies of ~85% being a reasonable goal (Pandit et al, 2009; Joshi, 2008).
Fig. 4. Plots of efficiency from Equation 1. Points for a list will lie on solid line when there are no cancellations. Where there is a cancellation, the line will be within this ‘triangle’ (eg, line for 50% patients cancelled is shown, dashed). Points A-E: explanation in text.

4. Measuring ‘productivity’ in addition to ‘efficiency’

One limitation of this notion of ‘efficiency’ is that it does not recognise actual work completed. Thus, where two teams work equally efficiently in performing, say, knee replacements (that is, they utilise equal proportions of their list-time in productive activity, without over-run or cancellation) it is still possible that one team completes 4 operations in 8 hours while the other completes only 3. We would of course like to conclude that the former is more ‘productive’ than the other, if they have otherwise worked equally efficiently. An acceptable measure of relative productivity in a scenario like this is simply the number of operations completed (‘operations per hour’). However, this cannot be a universal measure as teams rarely undertake just one operation and do not all work equally efficiently (one team may complete more cases but cancel more patients). ‘Operations per hour’ as a measure also biases in favour of shorter operations, while other measures such as ‘income’ favour operations – arbitrarily – priced highest (Abbott et al., 2011): and avoiding such biases enables comparison of teams across specialties.
If we were to develop a measure of ‘productivity’ (in addition to that of ‘efficiency’ above) that would suitably apply to all surgical lists, it would need to fulfil the following criteria, from first principles:

1. the measure should be independent of casemix. The procedures undertaken or the co-morbidities of patients should not influence whether a team is regarded as ‘productive’ or not. In other words, inherently short and long procedures should be regarded as potentially equally productive (speed in this sense being the time from the start of anaesthesia to the time of arrival of the patient in the recovery area);

2. for any given surgical procedure, productivity is inevitably related to the speed with which the operation is completed. For the same operation (eg, hernia repair) the faster team is reasonably regarded as the more productive;

3. however, adoption of new techniques which are inherently slower to achieve the same surgical aim should not result in a team being regarded as no longer productive. For example, laparoscopic techniques improve safety, pain scores or postoperative stay but can take longer to perform (Maione et al., 2005). A team which once completed, say, three open operations and now completes only two laparoscopically does not automatically make it less ‘productive’;

4. the greater the total anaesthetist-surgeon contact time with the patient during a list (ie, as a proportion of list-time), the greater should be the productivity measure for this list;

5. productivity should only be regarded as having increased when any time savings made by improved practices, reducing idle gaps or greater speed are used to accommodate extra cases, rather than finish the list early;

6. any measure ‘productivity’ should be applied only to lists that are acceptably ‘efficient’ (by Equation 1). Or expressed another way, any measure of ‘productivity’ should incorporate the measure of efficiency; if the list is inefficient, this should be reflected proportionately in a reduced measure of productivity.

It is possible to reflect these six sentiments in an empirical mathematical formula (Pandit et al. 2009):

**Equation 2:**

\[
\text{Productivity} = 111 \times [\text{Efficiency} - (10^{10} - (\text{Speed} \times \text{Patient Contact}))]
\]

Where ‘efficiency’ is calculated from Equation 1. ‘Speed’ is simply the relative speed of completing the surgical procedure. It is assigned any value, where 1.0 indicates average speed while, for example, 2.0 indicates working twice as fast. It is calculated by reference to published values or the team’s or hospital’s own average speed, or is assigned an arbitrary value of 1.0 if speed is unknown. ‘Patient contact’ is the proportion of list time spent in productive anaesthetic or surgical activity (the converse of this is ‘gap’ or ‘turnover time’). It can have any value from 0 (whole list wasted) to 1.0 (no gaps at all). Equation 2 yields a value for ‘productivity’ ranging from 0 to 100% (the last is attained when efficiency is 100%, speed is 1.0 and patient contact is 1.0).

Graphically, this formula can be plotted (Figure 5). Here, a list lies on its specific efficiency curve and its place depends on the product of speed and patient contact. Thus list A is ~75% efficient; list B is a little more efficient but has a little less patient contact and/or speed of surgery. List C has the same amount of patient contact and/or speed as A, but is far less efficient (eg, perhaps through over-runs and/or patient cancellation). List D has similarly poor efficiency as C, but even worse patient contact and/or speed. List E has the same
efficiency as A, but poorer patient contact and/or speed: E can increase its efficiency (eg, by reducing patient cancellations) and so move in the direction marked by the arrow.

Fig. 5. Productivity calculated by Equation 2 plotted against the product of speed and patient contact for lists A-E. Also shown are two lines of efficiency.

Although these relationships are empirical, they have been usefully modelled (Pandit et al., 2009). The six criteria which a measure of productivity should fulfil (listed above) have some analogy with measuring productivity in businesses that undertake skilled, complex tasks such as antique clock repair, as opposed to low skilled, repetitive tasks (Schmenner, 2004). High productivity results when the business accurately estimates its workload. That is, only accepting enough clocks for repair that occupy its capacity without overwhelming it and cancelling orders or postponing work. This is akin to sensible booking of patients onto operating lists (see below). Staff in the business should spend as much of their time working on the clocks, rather than in idle gaps or breaks (a notion akin to maximising patient contact on a surgical list). Finally, for any given clock, staff should ideally take no longer than the average time in repair for the complexity of the task. This is akin to the notion of speed.

We have assumed throughout that quality of service is maintained in all aspects of our desire for efficiency and productivity, as that is an essential aspect of service delivery. Without quality standards being met, there can be no true measures of efficiency/productivity at all (ie, a factory making televisions has zero productivity, regardless of how many it makes, if its televisions do not work).
5. Planning cases for a surgical list

If it is desirable to utilise a list as much as possible, but not over-run and to complete all the cases booked, it follows that effective list planning is an important aspect of a properly ‘lean organisation’. Is it possible to book a list rationally so that these aims are more easily met?

In many organisations, lists are booked directly by surgeons or their secretaries (or occasionally by managers) in an ad hoc manner, using their own experience to estimate whether the number and type of operations booked is appropriate for the time available on the list. However, they may face several pressures that cause them to over- or under-book the list. Surgeons may feel that over-booking demonstrates to others how hard they work, or their past surgical training may not have included organisational training, or surgeons may possess or develop character traits that make them prone to exceeding their own capacity for work or the presence of a large waiting list may be a worry. All these factors may cause them to over-fill lists. On the other hand, other demands on their time (teaching, lecturing, committee meetings) may prevent them from fully utilising a list. These issues have been discussed elsewhere (Jones & McCullough, 2007).

Fig. 6. Histogram of the number of surgical lists (of 150) with the degree of over- or under-run. The majority of lists over-ran, and only a minority (dark boxes) finished within 10% of the scheduled list time.

If patients scheduled for elective surgery first enter a ‘pool’ or waiting list (where they wait for several months; see http://www.nhs.uk/18weeks) then the problem appears to be a relatively simple one of ensuring a series of cases from this readily-available pool fills the pre-allocated block-time, with no over-running. Key to planning the list, therefore, should be knowledge of the average time (plus the standard deviation, SD) each case is likely to take.
The ad hoc method is distinctly poor at booking lists well (Figure 6). A more quantitative method can use the mean and SD of the published surgical procedure times (this referring to the time from start of anaesthesia to the arrival of the patient in the recovery area after the end of surgery) using the following equations (Pandit & Tavare, 2011):

Equation 3:

\[
\text{estimated mean duration of list} = (M_1 + M_2 + \cdots + M_x) + (G_t, S_t)
\]

Equation 4:

\[
pooled SD = \sqrt{\frac{SD_1^2 + SD_2^2 + \cdots + SD_x^2}{x}}
\]

M₁, M₂, etc refer to the mean times for the cases to be scheduled; Gₜ is the proportion of list time estimated to be wasted as gaps (usually ~10%, or 0.1) and Sₜ is the scheduled list time in min (e.g., 480 min for an 8 hour list); SD₁, SD₂, etc are the corresponding SDs for the respective operation times M₁, M₂ etc. The results of these equations, including the pooled SD can be used using the t-distribution to generate a probability that the proposed list will finish within the scheduled list time.

Fig. 7. For 150 surgical lists booked by the ad hoc method, the % actual over-run or under-run (y-axis) was plotted against the calculated probability that the list would over-run using Equations 2 and 3 and the t-distribution. Lists that suffered a cancellation are shown as hollow circles. Note that the ad hoc method yields a majority of lists that are predicted to over-run (and actually do so) or predicted to under-run (and actually do so).

Furthermore – and most importantly – the same t-distribution can be used to generate a probability that the proposed series of cases will exceed a certain minimum list time that should be fulfilled in order to utilise the list appropriately. In other words, if a list is scheduled for 480 min but it is judged that at least 450 min should be utilised, this forms a
lower boundary (B1): the probability estimated that the proposed list of cases exceeds this time should be relatively high (e.g., >80%). Yet, if it is also judged that the list should not over-run beyond 510 min then this becomes a higher boundary B2. The probability that the proposed list of cases exceeds B2 should be low (e.g., <20%). Taken together, therefore, B1 and B2 and the generated probabilities form a heuristic (a rule of thumb) that can be used to book lists (see: http://links.lww.com/EJA/A19).

Figure 7 shows that the ad hoc method of list booking is generating lists with probabilities of near-100% that they will over-run (many of these suffer a cancellation) and that many are booked so that they have a near 100% chance of under-running.

Equally, this same figure suggests that these probabilities can be used to book the lists in the first place. If this were done, Figure 8 shows that there would be good agreement with the predicted list time and the actual list time.

![Figure 8](http://links.lww.com/EJA/A19)

**Fig. 8.** The predicted list duration from Equations 3 and 4 (x-axis) fits the real data (y-axis) very well for the lists in Figure 6. Error bars are 95% confidence intervals of the estimate; the solid line is the line of best fit superimposed on the line of identity; dashed lines are 95% confidence intervals for the fit.

### 6. Conclusion

‘Lean’ can be adapted to the healthcare (surgical-anaesthetic) situation in a quantitative and tactical way, extending its application to just the operational level of management. This, as shown above, should result in the appropriate surgical capacity being provided for the workload, it should attain ‘efficiency’ and ‘productivity’ on the surgical list and it should facilitate proper scheduling of the cases on the list to achieve that efficiency.
7. References


www.intechopen.com


Pandit, J.J. & Tavare, A. (2011). Using mean duration and variation of procedure times to plan a list of surgical operations to fit into the scheduled list time. European Journal of Anaesthesiology, [In Press].


In the new millennium the increasing expectation of customers and products complexity has forced companies to find new solutions and better alternatives to improve the quality of their products. Lean and Six Sigma methodology provides the best solutions to many problems and can be used as an accelerator in industry, business and even health care sectors. Due to its flexible nature, the Lean and Six Sigma methodology was rapidly adopted by many top and even small companies. This book provides the necessary guidance for selecting, performing and evaluating various procedures of Lean and Six Sigma. In the book you will find personal experiences in the field of Lean and Six Sigma projects in business, industry and health sectors.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:
