

Report to the Bristol Royal Infirmary Inquiry

Learning curves in relation to surgery

A Discussion Paper

Professor Steve Gallivan

September 2000

Paper No. 572



Clinical Operational Research Unit
Department of Mathematics
University College London

1. Introduction
2. Background
3. Caveats about the notion of learning curves
4. Measurements of achievement
 - continuous
 - discrete
5. Assessment of learning dynamics
 - 5.1 Measures of performance
 - 5.2 Statistical methods for assessing learning effects
 - 5.3 Cusum methods – basic principles
 - 5.4 Practical examples
6. Non-Statistical issues associated with surgical learning
7. References

1. INTRODUCTION

Surgery requires a high degree of medical knowledge but is also highly dependent on manual dexterity. As with any task, it needs to be learned and often, but not always, the more practice one has the more proficient one becomes.

The term 'learning curve' is currently in vogue in relation to surgery. This rather jingoistic expression relates to a rather vague, and potentially misleading notion, that somehow the skill of a surgeon is acquired in some sort of continuous process. The concept implicitly assumes the notion that, if one had the means, there would be a quantitative measure of 'skill' which if displayed on a chart would show a continuous and elegant geometric curve describing the learning process. This is probably bunkum. For an individual surgeon, the 'curve', is more likely to be a jagged and irregular with at best with a trend indicating a move towards proficiency. Such is the nature of clinical data. Aggregating data from many surgeons may possibly produce a cloud of data points in which a underlying shape can be discerned, however, in most circumstances, one would expect an individual surgeon to show considerable divergence from this curve, so whether this provides a sensible means for monitoring performance is debatable.

Even if there were a platonically ideal geometric form describing the learning dynamics of a particular surgical procedure, for the most part we probably would never be able to gather sufficient data to characterise such a curve. Nor would one expect the same platonic shape to fit all surgical procedures.

2. BACKGROUND

The present author's experience of research into 'learning curves' and surgery is not great. However, it includes experience of 'growth curves', which has some similarities, and also research related to outcomes of cardiac surgery, both for adults^{1,2,3} and for children⁴. The author has written two articles that touch upon the subject of learning. One of these was a commentary on someone else's work (as yet unpublished) who was specifically discussing learning curves. In preparing this commentary, a brief literature review revealed that relatively little has been published on the topic. Thus while the author can not claim extensive expertise, it is probably comparable to most.

3. CAVEATS ABOUT THE NOTION OF LEARNING CURVES

There is little evidence that the notion of 'learning curve' is actually applicable to surgery, while plenty of evidence of other fields of human activity where it is patent nonsense, particularly in relation to activities such as golf and cricket, both notable for their reliance on dexterity. Even the most experienced wicket keepers are known to drop catches, and golfers drift in and out of form.

Undoubtedly it is true that there is a need to learn surgery and that junior surgeons with little experience are both slower and possibly less effective than their more

experienced colleagues. This in itself does not imply that there is a learning curve, just that there is learning.

For some elements related to learning, one either knows a thing or one does not, and there is no middle ground. For example, one either knows or does not know that wheel-nuts should be loosened before jacking up the car when changing a wheel. This knowledge is usually acquired all at once, usually the first time one has to do the job. Such knowledge does not accumulate in a gradual process over time related to the number of tyres one has changed. There are many surgical processes which have similar all-or-nothing characteristics: it is bad to slice through a major blood vessel, spurting blood should be investigated and clamped, double check which limb is to be amputated, etc. In such cases, the learning process is not characterised as a curve.

In spite of these potential difficulties with the concept of 'learning curves' in relation to surgery, there are many references to the phrase, indeed it occurs frequently within the Inquiry's evidence.

4. MEASUREMENTS OF ACHIEVEMENT

Assuming one wishes to use a quantitative method to assess how well a surgeon has 'learned' a particular procedure, there is the issue of what measurements to use. The fact is that there is no direct objective measurement available that can be used to assess a surgeon's state of skill. Instead surrogate measures are required. These are quantities that are believed to indicate indirectly how proficient a surgeon has become. These need to be treated with caution since the same factor that indicates proficiency, say operating speed, may also, possibly indicate poor practice if taken to extremes.

There are two broad classes of surrogate measure:

Continuous measures:

There are many examples of such measures. For example

- blood loss during operation;
- survival time following operation;
- operation time;
- time patient on heart/lung machine;
- time required in post-operative intensive care.

Discrete measures

For example

- Peri-operative mortality;
- post operative infection;
- 'near misses';
- brain damage;
- vaginal tearing during forceps delivery.

5. ASSESSMENT OF LEARNING DYNAMICS

Assuming quantitative data have been gathered about such factors for successive operations carried out by an individual surgeon, or indeed for a group of surgeons. If there are sufficient data about a single type of surgical procedure, one might consider investigating learning dynamics.

5.1 Measures of performance

There are many statistical options available depending whether one focuses on continuous or discrete measures:

Continuous measures

There appear to be several options available using various degrees of statistical sophistication. For example:

- compare the first half of a sequence of operations with the second. Assess the difference in mean or median 'performance' using t-test or Mann-Whitney test;
- use successive quartiles of data and construct box-and-whisker plots;
- examine non-linear fits to the data using a variety of types of curve which have 'logistic curve' characteristics. Such techniques often work well if data are logarithmically transformed prior to curve fitting, (although technically, error structures should be examined to confirm that this is appropriate, which is often not done).

The former methods have disadvantages in that they are somewhat 'arbitrary'. Also dividing data into arbitrary sections might miss the crucial period when 'learning' occurred, as the wheel changing example illustrates. If 'learning' occurs very early in a sequence of operations, and if the sequence is long, neither method would detect it.

Data fitting might be preferred, but this relies on considerable amounts of data being available. There are also considerable technical difficulties dealing with serial data from groups of surgeons, since such data are not statistically independent and thus usual curve fitting methods are invalid. Most important, one needs to have some idea of the mathematical nature of the underlying curve before one can sensibly do curve fitting. While statistical packages may allow one to fit a logistic curve to data that actually follow a discontinuous form, this doesn't make such analysis sensible.

If one is dealing with a new operation, or indeed just an infrequent operation, such as arterial switch, then such methods are certainly not applicable. Not only is it uncertain what curve to fit, but there are not enough data to fit a curve. Indeed, there are barely enough data to estimate mortality with precision⁴, let alone detect curvilinear trends in a single surgeon's performance.

Discrete measures of outcome

Here the statistical problems are rather more difficult, and this is very much the context of the Bristol Inquiry. The problem is that with all-or-nothing measures, such as peri-operative mortality, the outcome measure for a single operation gives only one usable piece of information (alive/dead). From a statistical viewpoint, the occurrence of deaths in a small number of operations does not represent very much evidence, particularly if mortality rates are high. There is only one piece of information for each operation.

This is very different from the case where a continuous outcome measure is used, which can be more informative. For example, if for two successive operations, a junior surgeon uses over 8 units of blood when the average for the operation is 3 units, then this might be taken as fairly strong evidence that there is a surgical deficiency. On the other hand, deaths following two successive operations might not be uncommon even for the most able surgeon if one is dealing with a high risk procedure – in cold statistical terms, sad as they are, these two events do not represent very much ‘information’.

5.2 Statistical methods for assessing learning effects

Although methods such as logistic regression might be used to assess how mortality depends on experience, the preference has been to aggregate data and use them to derive mortality rates, then use the mortality rates themselves as the ‘continuous’ outcome measure. The need to aggregate data in this way implicitly means that more data are needed for sensible analysis.

The processes of aggregating data and thus estimating mortality rates leads to methodological problems. Moving average methods give a useful graphical technique, but are difficult to interpret from a ‘formal’ hypothesis testing viewpoint. Indeed it is by no means clear whether formal hypothesis testing methodology is applicable in such circumstance, since there are so many uncontrolled factors.

Although some statisticians advocate rather sophisticated analytical methods for the analysis of mortality data, by and large these are neither appreciated nor commonly used by the clinical community.

5.3 Cusum methods – basic principles

So called Cusum methods seem to be the most popular means of examining learning effects in relation to mortality⁵. At their simplest, these show a cumulative running total of total mortality plotted against total number of operations performed. This gives a jagged curve that climbs upwards. This is illustrated in Figure 1. Learning effects can in principle be detected by observing how the slope of the Cusum curve changes. However, if the case mix of a surgeon changes, which is likely for a surgeon in training, then apparently stability of the cusum chart may shield the fact that the surgeon is dealing with more and more difficult cases.

More sophisticated analysis is available in the case where pre-operative risk factors are known, and estimates are available for the risks of death, which are possibly different for every patient. Here one can take the surgeon's case mix into account and derive an estimate for the number of deaths that might be expected from the individual risk forecasts. This is illustrated in Figure 2.

As shown in Figure 2, the difference between the curve expressing the expected number of deaths and the actual number, gives an estimate for the 'Net Life Gain'. Plotting this against the number of operations that a surgeon has done gives an informative summary of overall performance, as illustrated in Figure 3.

As illustrated in Figure 3, this gives a jagged curve that moves up, for each surviving patient, and dips for each death. The amount of rise or fall of the chart depends on the pre-operative risk of the patient considered. The higher the risk, the less the surgeon is 'penalised' if there is a death; equally, the higher the 'reward' if the patient survives.

The case illustrated in Figure 3, shows a chart typical of a surgeon who is performing consistently better than expected, at least in comparison with predicted mortality. In practice, the process of pre-operative risk estimation for adult cardiac surgery often makes use of so called Parsonnet estimates⁶, which tend to overestimate risks. As a result, if Parsonnet estimation is used, then net life gain estimates for experienced surgeons usually tend to drift upwards, as shown in Figure 3. Even if there is some overestimation, such curves can be used to detect changes in performance which correspond to changes in slope. Perhaps more important, such charts give a way of comparing performance.

This method is becoming a frequently used method for monitoring outcomes in relation to adult cardiac surgery. The first time this method was used for adult cardiac surgery was for a study conducted at St George's Hospital in London¹. For arcane reasons, the method became known by an acronym VLAD, although use of this name is common, it is not universal. Although a very simple graphical method for monitoring mortality, it was found in practice to be exceptionally good at alerting people to potential problems. VLAD was used as part of the analysis carried out for the College of Surgeon's investigations into Bristol (with striking effect) to complement the analysis carried out using more standard statistical techniques. The use of VLAD plots to compare (fictitious) surgeons is illustrated in Figure 4.

The VLAD method has since been extended to allow one to assess the likelihood that a departure from the norm could have occurred as a chance coincidence. Typical output from this extended version is illustrated in Figure 5.

In figure 5, the VLAD chart indicates the usual jagged profile of 'net life gain', in this case using real data. The display is supplemented by a bell shaped curve to the right which indicates the probability that a departure from the horizontal axis could have resulted purely as a chance coincidence. In this case, there is evidence of better than expected performance, although the evidence is not strong. Such apparent performance may well have been a fortunate coincidence. It should be stressed that such methods do not represent 'formal' statistical hypothesis testing.

There are certainly examples where the use of such VLAD methods have shown clear evidence of the 'learning' effect. However, the use of such methods is probably more

use in the monitoring of performance after the learning phase, as a means of detecting potential problems. Certainly, there are more sophisticated statistical methods becoming available, but as yet they do not seem to have not made a large impact on clinicians.

5.4 Practical examples

The examples given above have all been based on fictitious data chosen to illustrate the principles underlying Cusum methods. It is instructive to discuss some real life examples to illustrate how such methods are applied in practice. The examples in this section all concern adult cardiac surgery. In each case, a VLAD plot is used based on pre-operative risk forecasts made using the Parsonnet method. As already stressed, this tends to overestimate risks, so charts should in principle climb, a chart that hovers on or around the horizontal axis, or levels out for a considerable period indicates relatively poor performance rather than performance that is 'par for the course'.

The first example is taken from the original paper describing the VLAD method¹. Figure 6 shows a VLAD plot for a surgeon in training. This gives a rather stark illustration of the potential non-curvilinear nature of the learning 'curve'. The chart initially centres very much around the horizontal axis, relatively poor performance, however there is a sudden change, for whatever reason, and the chart then rises in a consistent fashion. Whether this discontinuity represents learning or some other factor that changed can not be determined from the data alone.

The second example, Figure 7, is from the same paper¹ and shows a VLAD plot for another surgeon and displays data for a longer sequence of operations. This too exhibits an initial period when the chart remains level and an apparent discontinuity when it starts to rise. Indeed, if only the initial part of the chart were shown, one might think the curve had similar characteristics to that of Figure 6 and that this initially period represented the 'learning' effect. However the surgeon concerned is very experienced surgeon, indeed a Professor of Cardio-thoracic surgery, so it is dubious to think of this as a learning curve. Also the initial operation in the sequence was chosen in an arbitrary fashion, merely because data for this sequence were readily available, rather than being all the operations that the surgeon had ever done. Caution is clearly needed when interpreting such charts. The chart is also instructive in that it illustrates several periods where the curve levels off. It is not possible to determine from the data whether these correspond to periods of true poor performance, or whether these are statistical 'blips', although the latter is not unlikely.

6. NON-STATISTICAL ISSUES ASSOCIATED WITH 'SURGICAL LEARNING'

The price paid for learning

There is clearly an ethical problem associated with surgical learning. For some procedures, for example the arterial switch, or even adult by-pass surgery, there is evidence that the less experienced surgeons have poorer results. Given this, there is an argument that only experienced surgeons should do such operations. However, if such a policy were followed precisely, there would never be the opportunity for more junior surgeons to gain the experience they need. Taken to extremes, a time would come

when all the experienced surgeons would have retired and the remaining surgeons would have to take on case load they have never acquired the necessary experience to deal with.

Inevitably, there are times that less experienced surgeons have to operate in order to learn their craft. Unfortunately, their patients may possibly be at higher risk. Possibly because this is such an ethically difficult issue, little research seems to have been carried to examine how this process operates in practice or how it could be improved.

Risk compensation and risk migration

In considering how to monitor surgical learning, it is important to recognise that there are many complicating factors. For example, in the case of cardiac surgery, junior surgeons do not start their training with the difficult cases, nor do they start their training unsupervised. Their caseload is deliberately skewed towards more straightforward cases and they are closely monitored by an experienced surgeon until they have gained experience. In the initial stages of training, should difficulties arise, the experienced surgeon will take over the operation. Such close supervision is maintained until it is judged that they have gained sufficient experience. Naturally, as experience is gained then, more and more difficult cases will be assigned to them. A consequence of this is that no 'learning' effects may be apparent from examination of Cusum charts, but would only be apparent from a deeper analysis of case mix and of 'rescues'.

An interesting notion is the theoretical possibility of risk migration. If case load requirements increase, the only option for the senior surgeons is to transfer some of their case load to more junior colleagues. Usually, they will transfer the cases that are the most straightforward, and have lower risk. As a consequence, the mortality for the senior surgeons would be expected to increase (since only the high risk cases remain). On the other hand, the cases transferred to the junior surgeons tend to be more difficult than they would be used, so as a consequence, mortality for the junior surgeons can also be expected to increase. Again, little research seems to have been done related to this rather paradoxical issue.

7. REFERENCES

1. J. Lovegrove, O. Valencia, T. Treasure, C. Sherlaw-Johnson, S. Gallivan, 'Monitoring the result of cardiac surgery by variable life adjusted display (VLAD)'
Lancet; **350**:1128-1130
2. J. Lovegrove, C. Sherlaw-Johnson, S. Gallivan, 'Monitoring the performance of cardiac surgeons'
Journal of the Operational Research Society; 50 (Number 7): 684-689, 1998.
3. C. Sherlaw-Johnson, J. Lovegrove, T. Treasure, S. Gallivan, 'Likely variations in perioperative mortality associated with cardiac surgery: when does high mortality reflect bad practice?'
Heart **84**:79-82, 2000
4. J. Stark, S. Gallivan, J. Lovegrove, J.R.L. Hamilton, J.L. Monro, J.C.S. Pollock, K.G. Watterson, 'Mortality rates after surgery for congenital heart defects in children and surgeons' performance',
Lancet **355**:1004-1007, 2000
5. de Leval MR, Francois K, Bull C, Brawn W, Spiegelhalter D. 'Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations'
The Journal of Thoracic and Cardiovascular Surgery, 107(3): 914 – 924, 1994.
6. Parsonnet V, Dean D, Berstein AD, 'A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease.'
Circulation (supplement I) ; 779(6): I-3 - I-12, 1989.