International Journal for Quality in Health Care 2000; **Volume 12, Number 5**: pp. 433–438

Methodology Matters

The use of the Cusum Technique in the assessment of trainee competence in new procedures

STEVE BOLSIN AND MARK COLSON

Department of Anaesthesia, Pain Management and Perioperative Medicine, Barwon Health, The Geelong Hospital, Victoria, Australia

Abstract

Continuous quality assurance (QA) in health care has necessitated the adoption of statistical methods developed as industrial process monitoring techniques. One such statistical technique is the cumulative summation (Cusum) methodology, which can monitor continuously a production process and detect subtle deviations from a preset defined level of achievement. The method is practical, simple to apply, easy to introduce and has proved popular with trainees in some specialities. This article introduces the concepts of a sequential analysis, deals with the practical steps of setting up a data collection and monitoring performance for procedures in health care.

Keywords: assessment, competence, Cusum, failure analysis, outcome, performance monitoring, quality assurance, training

There is increasing emphasis on Quality Assurance (QA), performance monitoring and credentialling in the practice of medicine and the delivery of health care [1–3]. The introduction of these techniques has been transferred from their use in other industrial and managerial processes [4]. The Cumulative Summation (Cusum) technique is one such statistical method, which has been proposed as a useful application in the field of physician and surgeon training [5–7].

This article will deal briefly with the important features of the technique, outline areas where the technique is seen as a particular advance and examine use of the technique specifically in personal professional monitoring. Finally, a number of sample graphs are presented which use simulated data to illustrate the expected failure analysis pattern in a variety of scenarios.

Background

The Cusum technique is one of a series of statistical tests developed during World War II as quality control tests for munitions production lines. The series of techniques known collectively as sequential analyses were originally described by Wald [8]. The first detailed description of the Cusum technique appeared in 1954 and the title *Continuous Inspection Schemes* reflects the language of the day [9].

The need for sequential analyses arose from several extensions of statistical techniques. These included the recognized shortcomings of repeated statistical tests of significance and the difficulties associated with tests in which the sample number was unknown but also expanding potentially *ad infinitum* [8].

The requirement for sequential testing was to develop a mathematical model which allowed the observer to decide if a production process was 'in control' (i.e. producing items within a defined quality boundary) or had become 'out of control'. In statistical terms this is formulated as changes in probability density functions of independent random variables occurring after an event, which represent a deleterious change in performance of the system or individual [8].

Under these circumstances part of the application of the Cusum technique is to identify the need for the stopping rule (the need to suspend the process, which is now 'out of control') as well as to choose the definitions of the stopping rule. The latter involves defining the boundaries of the quality envelope [8]. In this case the medical trainers define an 'acceptable' level of performance and this is used to formulate the stopping rule, which is applied to suspend unacceptable performance in trainees and initiate retraining (see Appendix).

Address reprint requests to S. Bolsin, Department of Anaesthesia, Pain Management and Perioperative Medicine, Barwon Health, The Geelong Hospital, Ryrie Street, Geelong, Victoria, Australia. E-mail steveb@barwonhealth.org.au

Trainers' input requirements

Cusum analysis can greatly assist medical trainers in their assessment of the competence of trainees. However, the technique offers no panacea to this difficult problem. The trainer must define the parameters on which the Cusum calculations will be based and ideally this should include valid results from procedures that are the subject of the data collection. The trainer must state from the outset what is an acceptable and unacceptable failure rate for the procedure in question.

The trainer must also determine the probability of false-positive and false-negative errors that is acceptable. A false-positive or type 1 error would lead to the conclusion that the trainee's performance is 'out of control' when it is not; a false-negative, or type 2 error, would lead to the conclusion that the trainee is 'in control' when they are not. The relative cost of either intervention to bring the trainee 'under control', or the cost of allowing the trainee to remain 'out of control' will influence the trainer's definition of the limits to activate the stopping rule. Such calculations may require input from actuaries, indemnity organizations and risk managers to cost unnecessary retraining efforts (type 1 error) against adverse events (type 2 error).

The medical trainer is not expected to bear the burden of accurate parameter determination alone in the development of Cusum analysis. The strength of the technique is that it enables pooling of data for the benefit of all participants. For a medical speciality trainee, the logical vehicle for this co-ordination is the relevant speciality association or college. Furthermore, the colleges have most to gain from the data collection process because one of their primary functions is to ascertain the competence of their prospective graduates. Cusum analysis offers a much-needed objective mechanism to assist in this awkward process and is being trialled by the Australian and New Zealand College of Anaesthetists for the assessment of the performance of selected practical procedures in trainee registrars.

Presentation of results

The performance data is best presented in a graph. Two main formats are described. The first presentation format described (Figure 1) is that used by de Leval [5]. The graph is of the number of cumulative failures on the vertical (*y*) axis, against the attempt number on the horizontal (*x*) axis. Thus, a zero failure rate would result in a horizontal line, but a 100% failure rate would result in a 45° line through the axis. As the cumulative failure count can never go down, the graph will rise inexorably but does provide simple intuitive information about crude success or failure rates at defined procedure numbers such as 10, 50 or 100. The boundary formulae are provided in the Appendix and represent the acceptable failure rates at any particular number of attempts. The boundaries define the quality envelope within which performance is acceptable. Higher failure rates are unacceptable and will trigger retraining. However an acceptable failure rate for a first year trainee may not apply to a senior trainee and can be adjusted by the p_1 and p_0 terms included in the P and Q values of the acceptable and unacceptable boundary line formulae.

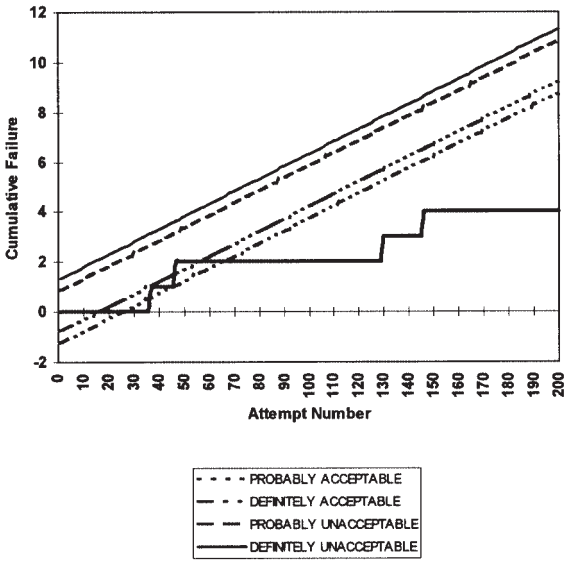

Figure 1 A Cumulative failure graph demonstrating performance of over 200 attempts at a procedure. A set of dotted lines denotes the tendency towards a boundary line, but to a lower standard of proof (α and β are set at 0.2 in this example). This demonstrates the performance to be probably acceptable by attempt 55, but we had to wait until attempt 66 to confirm acceptable performance. (Note that this is in exact accordance with the boundary intersection in Graph 2).

The second presentation format, used by Kestin, has the actual Cusum value plotted on the *y* axis against the attempt number on the *x* axis [6] (Figure 2). The Cusum value is the running sum of a mixture of increments (with each failure) and decrements (with each success), with the ratio between the two being determined according to the formulae outlined in the appendix. The decrease in the Cusum plot with each successful procedure completed is denoted 's' and the increase in the plot with each unsuccessful attempt is '1–s'. The value of s is related to the pre-defined acceptable and unacceptable failure rates. It follows that acceptable performance will be denoted on this format by a Cusum line which is roughly horizontal or down-sloping.

The Cusum formulae allow us to plot regular boundary lines that will embrace the defined parameters such as the type 1 and type 2 error rates, as well as the acceptable and the unacceptable failure rates. These horizontal lines are plotted at regular intervals on the *y* axis and are separated by values h_0 and h_1 but require some interpretation.

The Cusum for the series is plotted until it crosses either an acceptable boundary (from above) or an unacceptable boundary (from below). At that point it is possible to conclude that the performance during the preceding series of attempts

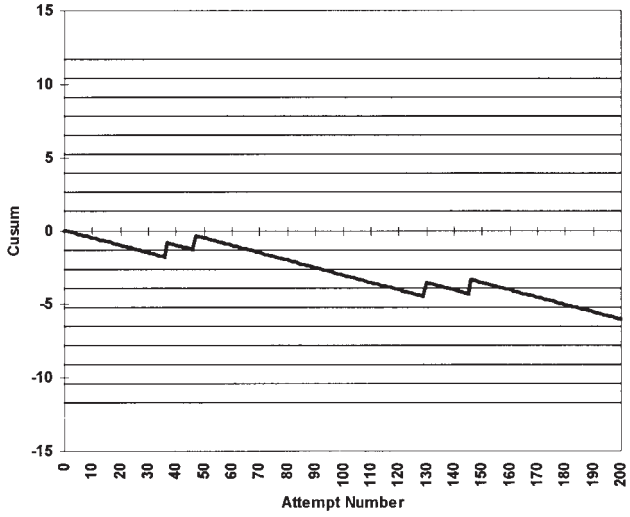

Figure 2 The same data as Graph 1 presented in the Cusum format. Failed attempts at the procedure are indicated by the upward deviations in the plot. The overall failure rate was four in 200 (2.0%). The acceptable failure rate was 2% and the unacceptable failure rate 10%. Type 1 and type 2 error rates were set at 0.1 to simplify the graph by making the spacing between acceptable and unacceptable boundaries identical. Competence of this operator is demonstrated by the fact that the Cusum plot spans four acceptable boundaries (in the downwards direction) but does not span any boundary lines in the upwards direction.

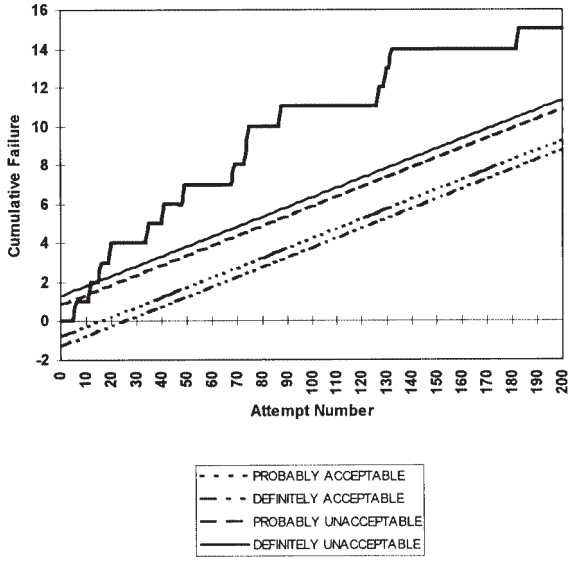

Figure 3 A typical learning curve in Cumulative failure format. Note that the performance improvement to an acceptable standard is by no means obvious in this format, demonstrating the weakness of this technique for long-term performance monitoring.

was either acceptable or unacceptable respectively, within the constraints of the entered criteria. Also one can re-start the analysis. Thus, if after intersecting an unacceptable boundary, the Cusum again rises to intersect another unacceptable boundary, it is possible to conclude that the performance during the series since the last boundary intersection has also been unacceptable. Likewise, if after intersecting an acceptable boundary, the Cusum again falls to intersect another acceptable boundary, then the performance has again been acceptable in the series of attempts since the previous boundary intersection.

The spacing of the unacceptable boundary lines is denoted h_0, while that of the acceptable boundary line is denoted h_1. The graph becomes unintelligible if both series of boundary lines are plotted in the positive and negative sectors. However, if we let the type 1 error rate equal the type 2 error rate, then h_0 and h_1 are equal and the lines become equally spaced. This is a major advantage since we then only need to plot one set of lines. In fact one set of boundary lines is superimposed on the other. This modest compromise eliminates the need to distinguish between the alternate types of boundaries – acceptable and unacceptable. Because a typical type 1 error is 0.05, while a typical type 2 error is 0.2, the logical choice for identical values of each is 0.1 [6].

While alternate presentation formats for the same data can cause confusion each format has certain advantages. Plotting the Cusum is ideal for long-term performance surveillance (as in continuous professional development), as one can readily identify a change in performance after a period of either acceptable or unacceptable performance. On a cumulative failure graph, such a change in performance is much more difficult to identify. It is also more difficult to determine the significance of any such change without re-plotting the data, from the first attempt of the series to be analysed. Nevertheless it is a suitable presentation format for small data sets.

The Cusum graph is admittedly a busy one that is only rendered intelligible by the compromise of allowing the type 1 and type 2 errors to be equal. The cumulative failure graph is not subject to this constraint as only one set of boundary lines is plotted. It is common practice to add a second set of 'alert' lines to the cumulative failure graph which use the same formulae, but a higher type 1 error (and usually type 2 error) value to alert the trainer to the fact that a trainee is approaching unacceptable performance. A suitable type 1 error value chosen for this purpose is 0.2 – in other words, a one in five chance of falsely accusing a trainee of unacceptable performance. These 'alert' lines are shown as dotted lines on the cumulative failure graphs (Figures 1 and 3). The activation of an alert state in a series of failures may enable early intervention, which may improve patient safety [5].

Applications

Cusum failure analysis lends itself to the surveillance of performance in virtually all aspects of procedural health care. Provided one can define strictly success and failure, and ensure the consistency of interpretation of its determination, a procedure lends itself to Cusum analysis. Consequently we

would propose that there exists in nursing, medicine and health management a vast area of performance monitoring which is currently neglected. The failure to apply a rigorous procedure to the analysis of success or failure in modern health care has inevitably resulted in numerous instances of unacceptable performance going 'unnoticed', and therefore, not acted upon [10]. The corollary of this is equally true: numerous clinicians must have lost confidence in their own abilities without good reason after a cluster of failures, which may be as low as two or three failures.

Training

The most difficult aspect of using Cusum analysis in training is determining what is an acceptable and an unacceptable failure rate. It is our belief that the acceptable failure rate should be the best estimate of the failure rate of a competent, experienced operator. The unacceptable failure rate is more difficult, but will typically lie in the range of two to five times the acceptable failure rate. As more Cusum data is collected appropriate success and failure rates will become defined for different groups by this performance data.

Obviously, new trainees who may be performing a procedure for the very first time will be likely to have unacceptable failure rates. There are two potential solutions for this: either one can adjust the failure rates for values which are appropriate for new trainees, or, alternatively, leave the failure rates unchanged and instead focus on whether performance ever becomes acceptable. The latter is the embodiment of a learning curve, and while it is a suitable approach for high-frequency procedures, it is less suitable for occasional procedures since in this case the trainee may take a disproportionately long time to ever finally demonstrate acceptable performance (Figures 3 and 4). In most clinical training settings, therefore, it is appropriate to judge trainees solely against the performance of their current and former colleagues with similar experience. The difficulty with this approach is that while failure rates for experienced operators abound in the literature, the corresponding data for trainees is scarce. An increased awareness by speciality associations and colleges of training performance should help to improve this dearth of information. In the meantime, individual departments can use their own local data to determine acceptable failure rates for trainees at each stage of their development (Figures 5 and 6).

Near misses

One of the potential problems with Cusum analysis is that it invariably focuses on hard end-points that are amenable to the determination of success or failure. A surgeon's in-hospital death rate is one such end-point [5]. While it is obviously immensely important, especially to prospective patients, it is likely to be a poor indicator of the surgeon's actual ability. The reason for this is that because death is

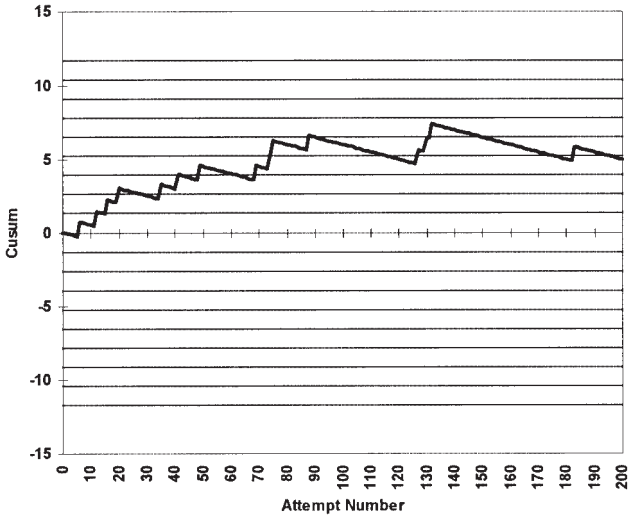

Figure 4 The same data as in Graph 3 in Cusum format. This operator begins with a failure rate of around 20%, thus spanning six unacceptable boundaries (from below). However, from attempt 90 onwards, the Cusum plot spans two acceptable boundaries (from above) and so his performance has improved to an acceptable level during the latter series consistent with effective training.

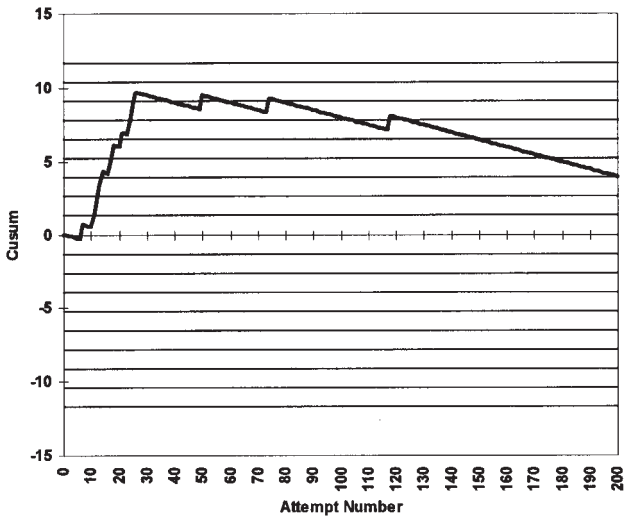

Figure 5 A disastrous learning curve with 'rescue' intervention. This trainee's poor performance was immediately evident and further attempts suspended when eight unacceptable boundaries had been crossed from below by the 30th attempt with a failure rate for this series of around 20%. The trainee was subject to intensive re-training and subsequently demonstrated acceptable performance with a failure rate for the next 170 attempts of around 2.0%.

such an undesirable outcome, hospitals are usually good at identifying patients who are likely to die, and enact numerous preventive strategies (often at great expense) to prevent further deterioration. For instance, the patient may be transferred to intensive care, and after weeks of sophisticated

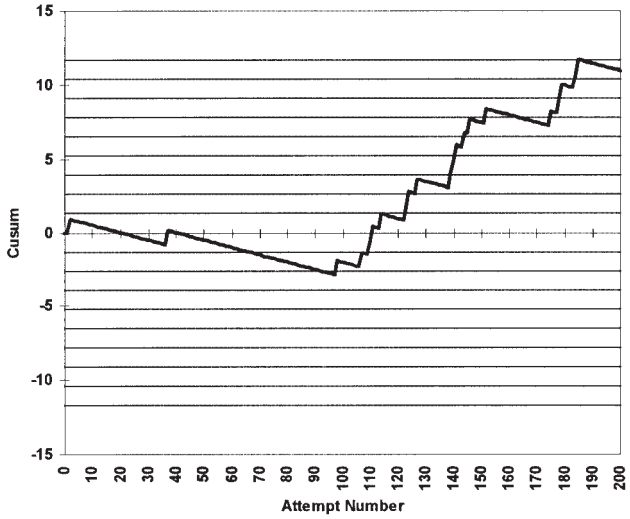

Figure 6 This operator demonstrates highly acceptable performance for the first 100 attempts at a procedure, during which time the failure rate is 2%. However, performance suddenly deteriorates (possibly due to disease) to become highly unacceptable, with a failure rate for the next 100 attempts of around 20%. This graph would identify the change in performance after as few as two failures had occurred.

treatment, eventually make a full recovery. Death is narrowly averted, and the surgeon can record another success.

A method of prospectively flagging potential adverse events and a solution to this apparent loophole is to add instances of 'near miss' to the analysis of any hard endpoint, such as death. Although 'near misses' are invariably much more difficult to define, the problem is not insurmountable [5]. For instance, we might define a surgical 'near-miss' as any case where the patient spent more than 1 week in intensive care after routine surgery. We might define an anaesthetist's 'near miss' for tracheal intubation as any instance in which he was successful but only with the assistance of another anaesthetist who had been urgently summoned.

'Near misses' can be graphed using either of the formats discussed above; however, in this case, the cumulative failure graph is probably more suitable. On this graph, the 'near-miss' plot will always be above (exceed) the adverse outcome plot. This technique was used by de Leval to demonstrate that in clinical practice selected near miss criteria can be used to predict deteriorating performance as they herald the occurrence of adverse incidents – in this case deaths [5]. If the gap between the two plots is reducing, the operator's ability to 'survive' a 'near miss' is reducing. Alternatively, if the gap is increasing, the operator (or at least the institution) is becoming more adept at averting the conversion of 'near misses' to adverse outcomes.

There are potentially enormous beneficial consequences for patient safety in the appropriate identification and monitoring of 'near miss' data that herald true adverse events in health care.

Appendix

Symbols used in formulae

p_0 = Acceptable failure rate
p_1 = Unacceptable failure rate
α = Type 1 failure rate
(The probability of wrongly accusing a trainee of unacceptable performance)
β = Type 2 failure rate
(The probability of wrongly certifying a trainee's performance to be acceptable)

Intermediate values

$a = \ln \{(1-\beta)/\alpha\}$
$b = \ln \{(1-\alpha)/\beta\}$
$P = \ln (p_1/p_0)$
$Q = \ln \{(1-p_0)/(1-p_1)\}$
Where ln is the natural logarithm ($\log_e$) of the function denoted
$s = Q/(P+Q)$
(s is the downward decrement with each success on a Cusum plot, while the upward increment with each failure is 1–s)

Cumulative failure graph formulae

N = Attempt number
$CF_{ACCEPTABLE} = sN - b/(P+Q)$
(Defines the boundary of acceptable performance on a cumulative failure graph)
$CF_{UNCCEPTABLE} = sN + a/(P+Q)$
(Defines the boundary of unacceptable performance on a cumulative failure graph)

Cusum graph formulae

$h_0 = b/(P+Q)$
(Defines the spacing between unacceptable boundary lines on a Cusum graph)
$h_1 = a/(P+Q)$
(Defines the spacing between acceptable boundary lines on a Cusum graph. Note that when $\alpha = \beta$, $h_0 = h_1$ and so the spacing between both sets of lines is the same)

References

1. Berwick DM. Continuous improvement as an ideal in health care. *New Engl J Med* 1989; **320:** 53–56.

2. Blumenthal D, Edwards JN. Involving physicians in total quality management: results of a study. In Blumenthal D, Scheck AC, eds, *Improving Clinical Practice*. San Francisco: Jossey-Bass, 1995: pp. 229–266.

3. Bolsin SN, Day CJ. Risk evaluation, quality of practice and audit. In Hall G, Morgan M, eds, *Short Practice of Anaesthesia*, First edn. London: Chapman Hall, 1998: pp. 111–122.

4. Shortell SM, Bennet CL, Byck GR. Assessing the impact of continuous quality improvement on clinical practice: what it

will take to accelerate progress. *Millbank Q* 1998; **76:** 593–624.

5. de Leval MR, Francois K, Bull C *et al.* Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994; **107:** 914–923.

6. Kestin IG. A statistical approach to measuring the competence of anaesthetic trainees at practical procedures. *Br J Anaesth* 1995; **75:** 805–809.

7. Williams SM, Parry BR, Schlup MMT. Quality control: an application of the Cusum. *BMJ* 1992; **304:** 1359–1361.

8. Siegmund D. *Sequential Analysis, Tests and Confidence Intervals*. New York: Springer, 1985: pp. 24–30.

9. Page ES. Continuous inspection schemes. *Biometrika* 1954; **41:** 100–115.

10. Bolsin SN. Education & Debate: The Bristol cardiac disaster. *BMJ* 1998; **317:** 1579–1580.