International Journal for Quality in Health Care 2002; **Volume 14, Number 3**: pp. 251–258

Assessing doctors' competence: application of CUSUM technique in monitoring doctors' performance

T. O. LIM, A. SORAYA, L. M. DING AND Z. MORAD

Clinical Research Centre, Kuala Lumpur Hospital, Ministry of Health, Malaysia

Abstract

Background. Quality assurance of medical practice requires assessment of doctors' performance, whether informally via a system such as peer review or more formally via one such as credentialing. Current methods of assessment are, however, subjective or implicit. More objective methods of assessment based on statistical process control technique such as cumulative sum (CUSUM) procedure may be helpful.

Objective. To determine the usefulness and acceptability of CUSUM charting for assessing doctors' performance.

Method. We applied CUSUM charting to assess doctors' performance of endoscopic retrograde pancreatography, renal and breast biopsies, thyroidectomy, and instrumental delivery. A CUSUM chart is a graphical representation of the trend in the outcome of a series of consecutive procedures. At acceptable levels of performance, the CUSUM curve is flat, while at unacceptable levels of performance, the curve slopes upward and eventually crosses a decision interval. When this occurs, the CUSUM chart indicates unsatisfactory performance. Thus, it provides an early warning of an adverse trend.

Results. All participating doctors found the technique useful to objectively measure their proficiency. CUSUM charts showed the progress of trainees in acquiring new skills. As they become more skilful with training, their CUSUM curves flatten. Among consultants, level CUSUM curves demonstrated ongoing maintenance of competence. All participants found the technique acceptable as a self-assessment tool. They were, however, less certain of its acceptability as a basis for credentialing.

Conclusion. We recommend the use of CUSUM charting as a tool for personal audit at an individual level. It may also be used to show proof of technical competence for the purpose of credentialing.

Keywords: clinical audit, competence, credentialing, CUSUM technique, outcome, performance monitoring, quality assurance, statistical process control

All countries need to ensure that the practice of medicine is ethical and competent, and thereby protect their public from poor practice. This is largely effected through a combination of legislation-like practitioner and hospital licensing laws, and professional self-regulation. Professional self-regulation itself is a privilege granted by the state through legislation. The privilege is typically vested in a national body, such as the General Medical Council in the UK and many Commonwealth countries. Self-regulation is essentially founded on the claim, among others, that the medical profession can be trusted to undertake the necessary action when individual doctors do not perform competently or ethically [1]. For self-regulation to be credible, the medical profession must demonstrate that it is capable of maintaining good practice. To that end, structures and processes to assure the quality of medical practice must be in place.

The quality assurance of medical practice in most countries is effected through a mixture of informal assessment and peer review, and more formal accreditation, credentialing or privilege delineation. The process of assessment, review, or credentialing is often subjective and without explicit reference to pre-determined standards of practice. It has been argued that comparative treatment outcome data on individual doctor's performance – so-called benchmarking – is required to make self-regulation credible [2]. Equally, we would argue that objective and quantitative methods to monitor the quality of a doctor's performance based on treatment outcome data could be more widely applied and would lend credence to

Address reprint requests to T. O. Lim, Clinical Research Centre, First Floor, Main Block, Hospital Kuala Lumpur, 50586, Jalan Pahang, Kuala Lumpur, Malaysia. E-mail: limto@crc.gov.my

the quality assurance process. A statistical technique to do so – the cumulative sum (CUSUM) procedure – has recently made its appearance in the medical literature [3–8]. Previous applications [3–8] of CUSUM have concentrated on a single discipline. We apply the technique to determine its utility and acceptability to doctors from a wide variety of disciplines.

Methods

Doctors from five disciplines participated in this study. These disciplines were nephrologists, gastroenterologists, radiologists, endocrine surgeons, and obstetricians.

Procedures assessed and outcome measures

1. Two nephrologists (a trainee and a consultant) were assessed to determine their competence at performing renal biopsy. Successful renal biopsy was defined as 10 or more glomeruli in the tissue obtained, the usual number required by a histopathologist for adequate interpretation.
2. Three gastroenterologists were assessed to determine their competence at performing endoscopic retrograde pancreatography (ERCP). Successful ERCP was defined as cannulation of the sphincter of Oddi, as determined by contrast radiography.
3. One radiologist's performance of stereotaxic core needle breast biopsy of non-palpable lesion detected by mammography from her training period through to her subsequent appointment as consultant was assessed. Successful breast biopsy was defined as adequate tissue for interpretation as judged by the reporting pathologist.
4. Two endocrine surgeons (a trainee and a consultant) were assessed to determine their competence at performing thyroidectomy under local anaesthesia. Two outcome measures were used to assess performance of this operation. They were: time taken to complete the operation (skin to skin), and pain experienced by the patient at second post-operative day as determined by the visual analog scale (VAS: scores range from 0 to 10, with 0 indicating no pain and 10 severe pain).
5. Three trainee obstetricians were assessed to determine their competence at performing instrument delivery using either metallic or silicon vacuum. The outcome measure used to assess performance of this procedure was failed instrumentation defined as failure to deliver the baby as intended.

CUSUM charting

The outcomes of these five procedures were monitored by CUSUM charts [9]. A CUSUM chart is basically a graphical representation of the trend in the outcomes of a series of consecutive procedures performed over time. It is designed to quickly detect change in performance associated with an unacceptable rate of adverse outcome. At an acceptable level of performance, the CUSUM curve runs randomly at or above a horizontal line (no slope). However, when an individual is performing at an unacceptable level, the CUSUM curve slopes upward and will eventually cross a decision interval; these are horizontal lines drawn across a CUSUM chart (see below for further explanation). When this occurs, the CUSUM is said to 'signal', indicating unsatisfactory performance. Thus, it provides an early warning of an adverse trend. A competent consultant is expected to have a level CUSUM curve, indicating ongoing maintenance of competence. On the other hand, a trainee in the process of acquiring a new skill is expected to have a rising CUSUM curve, the so-called learning curve. The degree of the slope is a measure of his or her progress in mastering the new skill: the greater the slope, the slower the progress. When the curve eventually flattens (no slope), this indicates he or she has mastered the new skill.

Design for CUSUM charting

Before a CUSUM monitoring scheme for doctors' performance can be started, several design decisions have to be made. We provide a non-technical account of a CUSUM chart design below. For a complete technical treatment, refer to Hawkins and Olwell [9] or another statistical text on the subject.

The CUSUM chart is a plot of the CUSUM score versus the index number of a series of consecutive procedures. Mathematically, the CUSUM score is determined after performance of each consecutive procedure when the outcome measure is known as follows.

For a CUSUM monitoring scheme designed to detect adverse deviation from an acceptable level of performance (referred to henceforth as Upward CUSUM): at the start, CUSUM $C_0 = 0$; at the nth procedure, CUSUM $C_n = \max(0, C_{n-1} + X_n - k)$; and the sequence C_n signals an upward shift in mean (i.e. indicating unacceptable performance has occurred) if $C_n > h$; where:

1. X_n is the outcome measure for the nth procedure. X_n is 0 or 1 for a binary outcome measure (success versus failure of procedure) with 1 indicating failure. For a continuous outcome measure (duration of operation and post-operative pain score for thyroidectomy in this study), X_n is the outcome measure standardized to have zero mean and unit standard deviation (SD).
2. k is the reference value and is determined by the prespecified standard of performance for the procedure to be monitored. For the binary outcome measure, the standard of performance is defined in terms of the acceptable and unacceptable failure rates in performance of the procedure. For the continuous outcome measure, acceptable level of performance is defined by the mean and SD of the outcome measure for a competent operator, and unacceptable performance is then defined by the size of upward shift in the mean in SD units.
3. h is the decision interval. When the sequence C_n exceeds h, the CUSUM monitoring scheme is said to signal, indicating that an unacceptable level of performance has occurred. When this happens, the doctor being monitored is required to determine and correct the

cause of the poor performance. The CUSUM monitoring scheme is then restarted. Restart should theoretically be at 0, but one often restarts at h as the new X-axis, so that a rising CUSUM graph can be obtained to represent the learning curve that is typically seen for a trainee.

h is determined by specifying the in-control (IC) and out-of control (OC) average run length (ARL) of a CUSUM chart. The IC-ARL is the average number of consecutive procedures required for a CUSUM chart to cross a decision interval despite an individual performing at an acceptable level. This is analogous to a Type I (alpha) or false positive error in hypothesis testing. The design with the short IC-ARL (large Type I error) is prone to false alarm. The OC-ARL is the average number of procedures performed before the CUSUM chart signal during the period when an individual is performing at an unacceptable level. The OC-ARL is a measure of sensitivity and is analogous to power [1-Type II (beta) or 1-false negative error] in hypothesis testing. The design with the short OC-ARL (high power) will quickly detect poor performance. In general, we want a CUSUM monitoring scheme to have long runs before a false alarm (long IC-ARL or small Type I error) and short runs before the chart signals actual deterioration in performance (short OC-ARL or high power). Unfortunately these objectives conflict, so we have to trade-off between them. This is also analogous to the trade-offs between Type I and Type II errors in hypothesis testing. Thus, a desirably long IC-ARL (small Type I error) will lead to an unacceptably long OC-ARL (low power). On the other hand, the desired short OC-ARL (high power) will lead to more frequent false alarms (large Type I error). The amount of trade-off between IC- and OC-ARL that is acceptable to the doctor clearly depends on the nature of what is being monitored. For example, a monitoring scheme for cardiothoracic surgery that entails life-threatening complications would require a highly sensitive chart to detect poor performance but at the expense of more frequent false alarms. On the other hand, for a procedure like renal biopsy, we would be prepared to tolerate a less sensitive scheme so as not to be frequently distracted by false alarms.

The participating doctors in this study specify the acceptable IC- and OC-ARL for monitoring their performance. Once these are specified, the decision interval h can be calculated [9]. The larger the specified IC-ARL (the OC-ARL will be correspondingly large), the larger is h. We could have specified the CUSUM design in terms of Type I and Type II error rates since they are analogous to IC- and OC-ARL, respectively. However, in our experience in designing the various CUSUM monitoring schemes in this study, it turns out that specification in terms of ARL was more intuitive and easier to explain to doctors. This is important because their inputs are required when designing a CUSUM scheme. For example, not a single doctor participating in this study could understand what a beta of 0.2 (power = 0.8) means in relation to monitoring and how changes to beta could affect the scheme; while it was easy to explain to them that an OC-ARL of 12 means that for an operator performing at an unacceptable level, on the average the chart would take 12 consecutive procedures before it signals. If they find 12 unacceptable, they could suggest a higher or lower number that they may be more comfortable with before being subjected to monitoring. Specification in terms of ARL also makes explicit the trade-offs between sensitivity and false alarm, and forces participants to be aware of the trade-offs they are making when their inputs are sought at the design stage of the CUSUM scheme. Otherwise, as a result of lack of understanding, there is a tendency to resort to conventional specifications like power = 0.8 and alpha = 0.05, as in hypothesis testing in the context of clinical trial. It is obviously undesirable to have one set of specifications for all procedures being monitored. For the purpose of monitoring, trade-offs between alpha and beta error should be allowed to vary depending on the nature of the procedure being monitored.

4. max $(0, C_{n-1} + X_n - k)$ is the maximum function that returns the larger of the two arguments, 0 and $C_{n-1} + X_n - k$. This function applies only to monitoring for an upward shift in mean (upward CUSUM). That is, monitoring to detect deviation from an acceptable to an unacceptable level of performance. This was the purpose of this study. For a scheme designed to detect 'better' than acceptable performance, the function is min $(0, C_{n-1} + X_n - k)$ with a signal if $C_n < -h$. Such a scheme (downward CUSUM) is not defined for this study for several reasons:

Some acceptable standards are so good [for example the 2% failure rate for breast biopsy (see below)] that designing to detect better performance at say 1% is difficult.

There was genuinely no interest at all in detecting 'better' than acceptable performance. Acceptable performance ought to reflect the performance of trained and experienced operators. Admittedly, a few exceptional individuals may perform better than their peers. It is, however, undesirable to base a monitoring scheme on results of 'star' performers. On the other hand, if most experienced operators performed at the 'better' level, it is only logical to define that level as the acceptable level.

Upward CUSUM chart design for the procedures studied

In designing the CUSUM monitoring schemes used in this study, we have to explicitly specify the following for each procedure being monitored:

1. Acceptable and unacceptable levels of performance for the chosen outcome measure. Ideally these should be based on universally accepted standards published by authoritative medical professional bodies. Unfortunately, to our knowledge, such performance standards

are not available. Their absence, however, should not preclude local or national groups from determining their own standards for the purpose of monitoring or clinical audit. In this study, for ERCP an acceptable failure rate of 10% seems reasonable and a failure rate of 20% is unacceptable [10]. For renal biopsy, in the absence of guidance from the literature, consensus among local nephrologists suggests that acceptable and unacceptable failure rates are 10 and 20%, respectively. For stereotaxic core needle breast biopsy, experience from specialist centres doing large numbers of such procedures suggests that an acceptable failure rate is 2% and a 5% failure rate is unacceptable. For instrument delivery, again there is surprisingly no guidance at all from the literature or professional body. Consensus among local obstetricians suggests that acceptable and unacceptable failure rates are 6 and 12%, respectively, for failed instrumentation. For thyroidectomy under local anaesthesia, an acceptable mean and a standard deviation for the performance measures have not been published. Therefore, the consultant surgeon's performance was taken as the standard, since hardly anyone else in this country or worldwide performs thyroidectomy under local anaesthesia. The monitoring scheme for thyroidectomy was designed to detect an increase in 1 SD unit in both the performance measures from their respective standards.

2. IC- and OC-ARL. For renal biopsy and ERCP, the IC- and OC-ARL were 52 and 16, respectively; for breast biopsy, the corresponding ARLs were 175 and 52; for thyroidectomy 500 and 9; and for instrument delivery 59 and 22, respectively.

Results

Figure 1 shows the upward CUSUM chart of a consultant and a trainee nephrologist for a series of 47 and 43 renal biopsy procedures, respectively. The consultant's upward CUSUM curve is flat, indicating that his performance has met the specified standard for the procedure. He has demonstrated ongoing maintenance of competence in performing this procedure. In contrast, the trainee's upward CUSUM curve was rising initially. The CUSUM crosses the decision interval for the first time at the seventh biopsy, indicating failure to meet the specified standard of performance. The CUSUM rises further and again crosses the next decision interval after 12 procedures. He has failed again. Nevertheless, he is making progress from then on; his upward CUSUM curve appears to level off for the next 23 procedures. However, from the 34th procedure onwards, he had two failures in close succession, causing his CUSUM graph to cross the next decision interval line.

Figure 2 shows the upward CUSUM chart of three gastroenterologists, labeled as doctors A, B, and C. They performed 17, 30, and 54 ERCP procedures, respectively. Doctor C is clearly the most competent of the three doctors in performing ERCP. After the first 20 procedures, he has a level upward CUSUM curve indicating performance at the agreed standard. In contrast, the curves of doctors A and B continue to rise as a result of a long series of consecutive failures. Neither curve shows any tendency to flatten out. Doctors A and B are obviously still struggling to acquire a new skill after performing 17 and 30 procedures, respectively.

Figure 3 is a typical representation of the learning curve of a radiologist in mastering the skill of performing stereotaxic core needle breast biopsy over 43 consecutive procedures. The upward CUSUM curve was rising for the first 10 procedures, the learning phase for the radiologist. Thereafter, the curve flattens out; she is then clearly competent at performing breast biopsy to the agreed standard.

Figures 4 and 5 show the upward CUSUM charts of a consultant and a trainee endocrine surgeon for the two performance measures: duration of operation and postoperative pain score. The consultant and trainee performed 23 and 39 thyroidectomy procedures, respectively. The upward CUSUM chart for the consultant is level by design for we assume the current performance of the consultant represents the acceptable standard (that is, the consultant is competent by definition, and not with reference to an external standard that is undefined for this operation). Note that even the consultant surgeon's performance with respect to duration of the operation showed a small learning curve early on during the first few procedures he attempted. He took 10 procedures before his CUSUM curve leveled off. CUSUM charts of the trainee for both performance measures demonstrate the classic learning curve pattern.

Figure 6 shows the upward CUSUM charts of three trainee obstetricians, labeled as doctors A, B, and C. They performed 49, 26, and 26 instrument deliveries, respectively. The three trainees clearly demonstrated varying learning curves. Doctor C had no learning curve at all; his upward CUSUM curve was flat from the first procedure he attempted. Both doctor A and doctor B had a learning phase. However, doctor B took only 11 procedures before his upward CUSUM curve begins to level, while doctor A required 23 procedures to achieve the same proficiency.

Discussion

We have demonstrated the utility of upward CUSUM charting in monitoring the quality of performance of doctors from five disciplines in performing a variety of procedures. All participants in this study have found the upward CUSUM technique useful in helping them to measure their proficiency objectively. For consultants, the demonstration of ongoing maintenance of competence in performing a particular procedure was reassuring. For the two trained doctors in the study, their failure to achieve the agreed standard of performance was unexpected. For trainees, upward CUSUM charting was helpful in monitoring their progress in acquiring a new skill. All participants have similarly found the technique acceptable, particularly as a personal self-assessment tool. They were, however, less certain of its acceptability as a basis for credentialing.

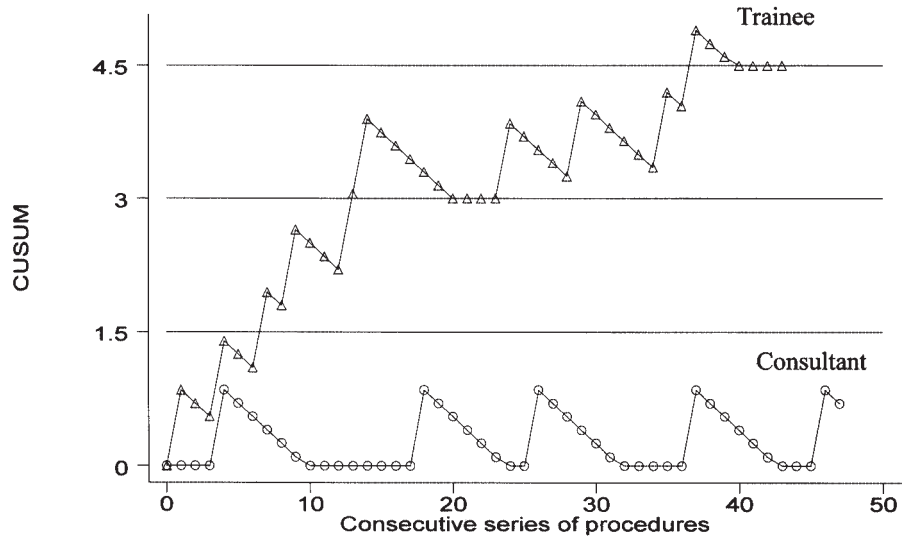

Figure 1 CUSUM for renal biopsy by a consultant and a trainee nephrologist.

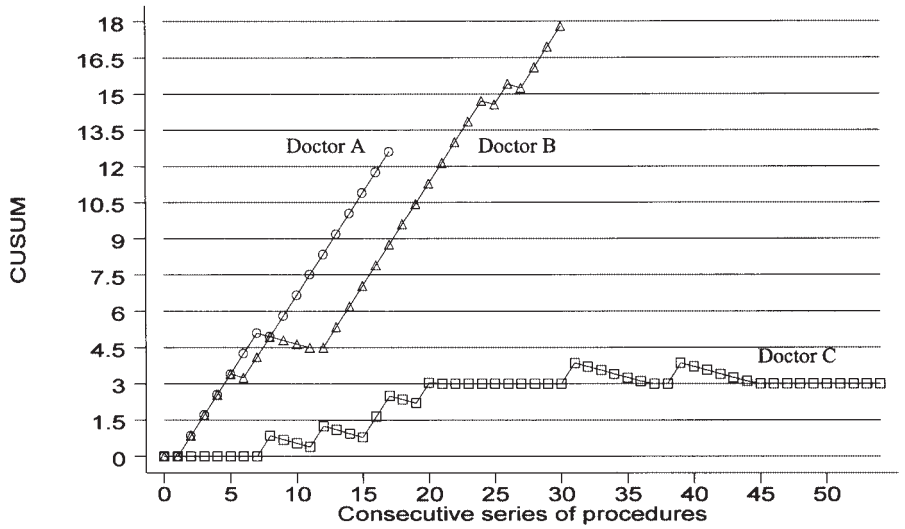

Figure 2 CUSUM for ERCP by three gastroenterologists.

Statistical process control (SPC) tools like the control charts have been widely used in the manufacturing industry for a long time. More recently it has seen its application in health care as well [11–15]. Although the use of SPC is well established in some disciplines like laboratory medicine [16], its application in clinical care processes poses special challenges. The best known control charts are those pioneered by Walter Shewart [17], for example his Xbar and R charts. Shewart charts, however, are designed to detect a large but transient shift in the process mean, typically in large-volume manufacturing processes. This limits their application in the clinical care process for two reasons. Firstly, the throughput of the clinical process is typically very slow; for example, a surgeon may perform no more than one to five procedures a day. It is both undesirable and inconvenient for a performance monitoring system to require sample sizes of greater than one to accumulate before analysis. Secondly, for clinical care, even a small shift in process mean is of concern; for example, adverse deterioration in mortality rate, complication rate, or procedure failure rate. Clinical monitoring requires early warning of poor performance before too many adverse outcomes have occurred.

The upward CUSUM chart is ideally suited for both of these requirements and has additional advantages:

1. It works for individual observation as well as for grouped observations (sample size greater than one).
2. It can be designed to detect a small shift in process mean. The CUSUM chart achieves its superior detection ability by accumulating information from successive

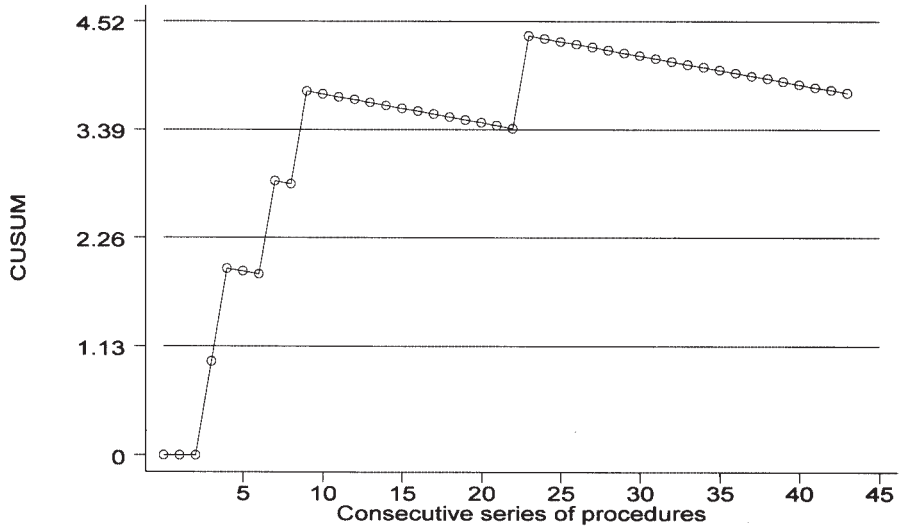

Figure 3 CUSUM for stereotaxic core needle breast biopsy by one radiologist.

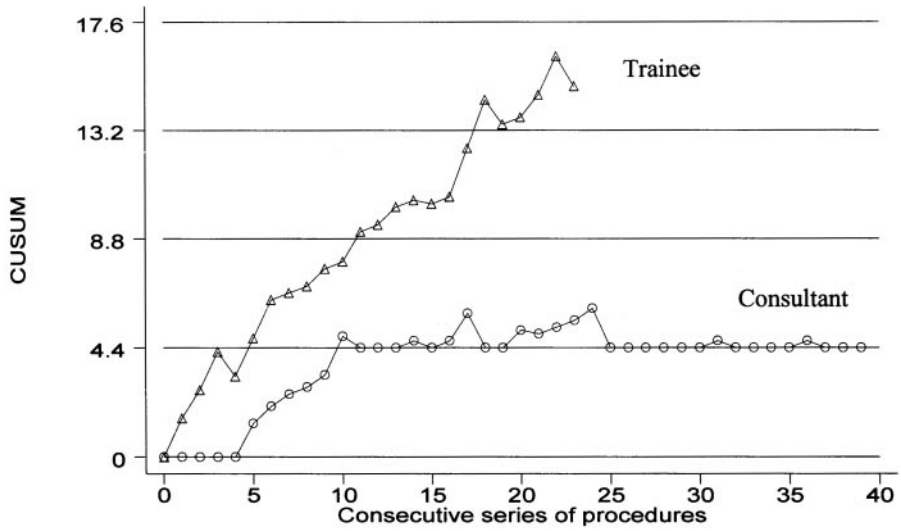

Figure 4 CUSUM for thyroidectomy under local anaesthesia by a consultant and a trainee endocrine surgeon. Performance measure is duration of operation.

deviations of a process performance from its targeted value. This allows a small difference to accumulate until a strong signal can be observed.

3. CUSUM charting requires specification of a targeted value for the outcome measure. This translates into requirements for explicit specification of a relevant outcome measure and a standard based on that outcome measure. An explicitly and unambiguously stated goal is desirable in quality assurance of medical practice. However, in this study, all groups of doctors had difficulty in setting quantitative standards. Published, universally accepted, and validated standards of performance for the procedures included in this study were not available.
4. CUSUM charting also makes explicit the trade-off between sensitivity and false alarm that is inherent in any monitoring system.
5. CUSUM charting is objective and has great visual appeal. This was a highly attractive feature to participants in this study. For trainees, it literally shows a learning curve and how an individual is making progress over time with more practice. This can complement the current system that relies on inspection by external observer, and is certainly better than relying on performance of an arbitrary number of procedures before competence is assumed [18–22]. As shown in this study, doctors do have varying learning curves. For consultants, CUSUM charting can clearly be used to

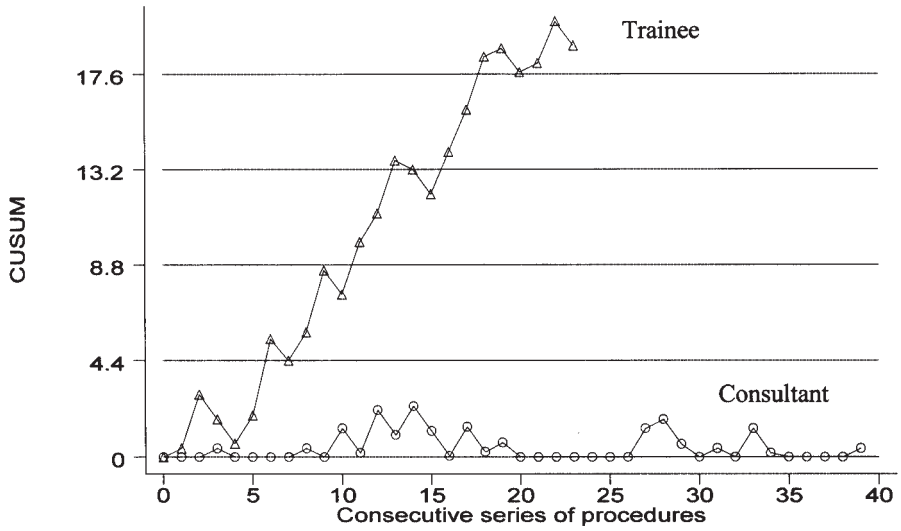

Figure 5 CUSUM for thyroidectomy under local anaesthesia by a consultant and a trainee endocrine surgeon. Performance measure is post-operative VAS pain score.

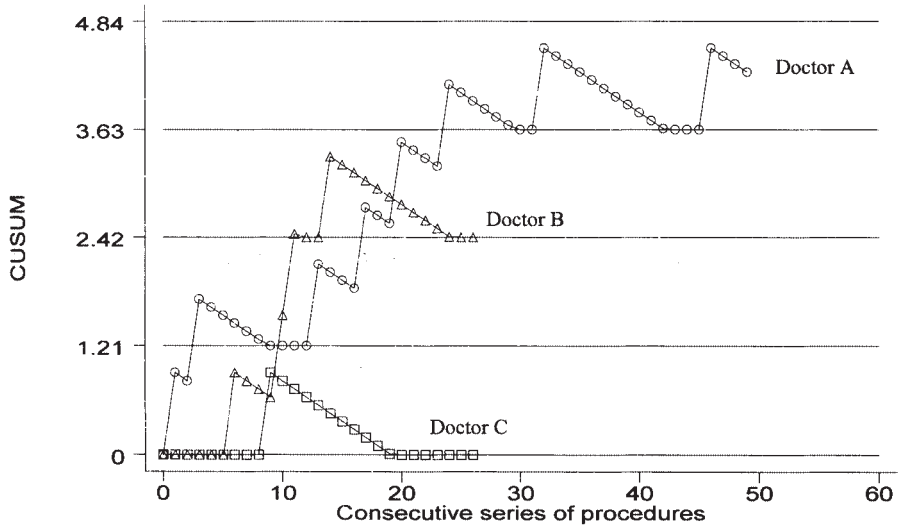

Figure 6 CUSUM for instrument delivery by three trainee obstetricians.

show proof of competence in a particular skill. Its objectivity can minimize the potential for bias in assessment, for example for the purpose of credentialing, and thus minimize the potential for conflict. This would also make the process of credentialing and quality assurance of practice in general more transparent.

To be credible, quality assurance of medical practice and professional self-regulation must incorporate elements of outcome assessment and peer review [1,2]. It is helpful to view this as a continuum, starting with the individual doctor who conducts personal assessment and extending through the clinical team and peer networks conducting peer review against a locally specified standard, through to national professional bodies that set national standards and conduct external review [1]. In this continuum, it is clear from this study that all clinicians found CUSUM most useful and acceptable as a tool for personal audit and perhaps also for local peer review. There was, however, considerable apprehension about its use at the national level by an external assessor or reviewer. In particular, how should one deal with suboptimal performance? For a trainee, this is no more than a manifestation of the learning process. When this occurs for a previously competent individual, a consultant for example, he or she is presumably under some obligation to examine his or her performance and to take corrective action, including retraining if necessary and, perhaps, his or her earlier privileges may have to be retracted. But this is an extremely sensitive issue and would require considerable fortitude. There are also issues about the confidentiality of

such information and the potential for medico-legal action. CUSUM charting has other limitations. It may be difficult to apply in some disciplines. It works best where performance has a quantitative outcome that can be measured reliably and in a timely manner, as for procedural disciplines. Its usefulness may be limited in cognitive disciplines. Furthermore, it assesses only technical skills. Other skills such as communication and interpersonal skills are no less important but cannot be assessed by CUSUM.

In conclusion, we recommend the use of the CUSUM chart as a method for assessing technical performance. It is a useful tool for personal audit at an individual and local level. At the national level, CUSUM chart may serve the need of a quantitative measure of proficiency for the purpose of credentialing but can be threatening.

Acknowledgements

The authors would like to thank the following persons whose work contributed to this paper. From the Kuala Lumpur Hospital we would like to thank S. S. Tan and S. T. Kew from the Department of Medicine, S. I. Yun and P. Sathyamoorthy from the Department of Radiology, and A. Noor Hisham from the Department of Endocrine and Breast Surgery. We also thank L. K. Soh and K. Y. Ng from the Department of Obstetrics and Gynaecology, Hospital Tuanku Ampuan Rahimah, Klang.

References

1. Irvine D. The performance of doctors I: professionalism and self regulation in a changing world. *Br Med J* 1997; **314:** 1540–1542.

2 Johnson JN. Making self regulation credible. *Br Med J* 1998; **316:** 1847–1848.

3 De Leval MR. Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994; **107:** 914–924.

3. Parry BR, Williams SM. Competency and the colonoscopist: a learning curve. *Aust NZ J Surg* 1991; **61:** 419–422.

5. Williams SM, Parry BR, Schlup MM. Quality control: an application of the CUSUM. *Br Med J* 1992; **304:** 1359–1361.

6. Van Rij AM, McDonald JR, Pettigrew RA *et al.* Cusum as an aid to early assessment of the surgical trainee. *Br J Surg* 1995; **82:** 1500–1503.

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

8. Bolsin S, Colson M. The use of the Cusum technique in the assessment of trainee competence in new procedures. *Int J Qual Health Care* 2000; **12:** 433–438.

9. Hawkins DM, Olwell DH. *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer-Verlag, 1998.

10. Cotton PB, Williams CB. *Practical GI Endoscopy*. London: Blackwell Scientific, 1982.

11. Atkinson S. Application of statistical process control in health care. *Managed Care Q* 1994; **2:** 57–69.

12. Shahian DM, Williamson WA, Svensson LG *et al.* Applications of SQC to cardiac surgery. *Ann Thorac Surg* 1996; **62:** 1351–1359.

13. Laffel G, Blumenthal D. The case for using industrial quality management in health care organizations. *J Am Med Assoc* 1989; **262:** 2869–2873.

14. Blumenthal D. Total quality management and physicians' clinical decisions. *J Am Med Assoc* 1993; **269:** 2775–2778.

15. Kritchevsky SB, Simmons BP. Continuous quality improvement: concepts and applications for physician care. *J Am Med Assoc* 1991; **266:** 1817–1823.

16. Garfield FM. *Quality Assurance Principles for Analytic Laboratories*, 2nd edn. Washington, DC: Association of Analytical Chemists, 1991.

17. Shewart WA. *Economic Control of Quality of Manufactured Product*. New York: Van Nostrand Reinhold Co., 1931.

18. Wilson FC. Credentialling in medicine. *Ann Thorac Surg* 1993; **55:** 1345–1348.

19. Benson JA, Cohen S. Evaluation of procedural skills in gastroentereologists. *Gastroenterology* 1987; **92:** 254–255.

20. Roberts JS. Privilege delineation in a demanding new environment. *Ann Intern Med* 1998; **108:** 880–886.

21. Hogan WJ. What constitutes endoscopic competence? *Gastroenterology* 1993; **104:** 1564–1565.

22. Greganti A, McGaghie WC, Mattern WD. Toward consensus: training in procedural skills for internal medicine residents. *Arch Intern Med* 1984; **144:** 1177–1179.

Accepted for publication 25 January 2002