Original article

Quantitative and individualized assessment of the learning curve using LC-CUSUM

D. J. Biau[1], S. M. Williams[3], M. M. Schlup[4], R. S. Nizard[2] and R. Porcher[1]

[1]Département de Biostatistique et Informatique Médicale, Assistance Publique – Hôpitaux de Paris (AP-HP), Hôpital Saint-Louis, Université Paris 7, Institut National de la Santé et de la Recherche Médicale U717, and [2]Département de Chirurgie Orthopédique, AP-HP, Hôpital Lariboisière, Université Paris 7, Paris, France, and Departments of [3]Preventive and Social Medicine and [4]Medical and Surgical Sciences, Dunedin School of Medicine, University of Otago, Dunedin, New Zealand

Correspondence to: Dr D. J. Biau, Département de Biostatistique et Informatique Médicale, Assistance Publique – Hôpitaux de Paris, Hôpital Saint-Louis, Université Paris 7, Institut National de la Santé et de la Recherche Médicale U717, 1 Avenue Claude Vellefaux, 75010, Paris, France (e-mail: djmbiau@yahoo.fr)

Background: **Current methods available for assessing the learning curve, such as a predefined number of procedures or direct observation by a tutor, are unsatisfactory. A new tool, the cumulative summation test for learning curve (LC-CUSUM), has been developed that allows quantitative and individual assessment of the learning curve.**

Methods: **Some 532 endoscopic retrograde cholangiopancreatographies (ERCPs) performed by one endoscopist over 8 years were analysed retrospectively using LC-CUSUM to assess the learning curve. The procedure was new to the endoscopist and monitored prospectively in the initial study. Success of the procedure was defined as cannulation and proper visualization of the duct(s) selected before the examination.**

Results: **Fifty ERCPs were considered unsuccessful. There was a gradual improvement in performance over time from a success rate of 82·0 per cent for the first 100 procedures to 96·1 per cent for the last 129 procedures. The LC-CUSUM signalled at the 79th procedure, indicating that sufficient evidence had accumulated to prove that the endoscopist was competent.**

Conclusion: **LC-CUSUM allows quantitative monitoring of individual performance during the learning process.**

Paper accepted 8 May 2008

Published online 22 May 2008 in Wiley InterScience (www.bjs.co.uk). **DOI:** 10.1002/bjs.6056

Introduction

Quality control in medicine has generated considerable interest in the past decade from public health authorities, doctors and patients[1–4]. More recently, quality control procedures have been applied to the assessment of learning curves of trainees and used to monitor the introduction of innovative technologies[5,6].

Various methods can be used to determine whether a trainee has reached proficiency and supervision can stop. Performing a recommended number of procedures is standard practice. This is not, however, tailored to the individual and some trainees may not have reached the required level of competence when the time comes for the first unsupervised procedure[7]. Direct observation by a tutor[8] and graphical representation of the learning curve, by use of a cumulative summation (CUSUM) graph[9,10], may be used to assess individual surgical performance. These qualitative methods, however, have no formal means of indicating when the required level of competence has been reached, and their objectivity and reproducibility is questionable[11]. The CUSUM test has been applied to the learning curve to address this problem[12]. It is designed to indicate when a process deviates from an acceptable level of performance[13,14] and is not necessarily suitable for monitoring the learning curve[15]. This paper describes the development of the CUSUM test for learning curve (LC-CUSUM), a quantitative and individualized statistical tool that may help determine when the learning curve for a surgical procedure is complete.

Methods

A systematic literature search of Medline via PubMed was conducted with the search terms '(CUSUM OR cumulative sums) AND (surgery OR procedure)' on 15 March 2006 supplemented by cross-checking of reference lists. There were no restrictions concerning date of publication, language or publication status. Twenty-two studies that assessed the learning curve were identified, of which one was selected because of the type of outcome presented (binomial), the number of procedures monitored (more than 70) and the shape of the learning curve. The authors were contacted and asked to participate in the present study; this analysis was based on the data retrieved.

This study evaluated the results of endoscopic retrograde cholangiopancreatography (ERCP) for biliary tract disorders[9]. ERCP considerably reduces the need for surgery, making it a cost-effective procedure. All procedures were performed by one endoscopist with several years' experience in endoscopy but with previous practical experience in ERCP limited to an introductory course. During an 8-year period, beginning in 1986, 532 ERCPs were performed. Relevant information such as indication, results, duct selection, cannulation and intervention were recorded for all patients. Data were obtained retrospectively from the patients' hospital notes and ERCP reports for first 2 years, and prospectively at the time of examination for the rest of the study.

Success of the procedure was defined as cannulation and proper visualization of the duct(s) selected before the examination. Duct selection was based on the patient's symptoms and results of pre-ERCP investigation. Inadvertent cannulation of the non-selected duct was not considered a failure provided that the selected duct was cannulated as well. More details may be found in the original report[9].

Statistical analysis

The standard CUSUM test was designed to monitor a sequential procedure (for example, a run of interventions), with ability to reject the null hypothesis, H_0, that the process is in control[13]. The alternative hypothesis, H_1, is that the process is out of control. Mathematically, a CUSUM analysis involves plotting the following quantity against the number of procedures, t:

$$S_t = \max(0, S_{t-1} + W_t)$$

with initial value $S_0 = 0$ and where W_t is a measure of the deviation of the outcome from the target. The process is assumed to be acceptable as long as the CUSUM score remains below a limit, known as h. The LC-CUSUM was developed to determine whether a process has reached a predefined level of performance. It presumes that the process is not in control at the start of monitoring (the trainee is not proficient) and signals when the process can be considered to be in control, in other words that the trainee has reached the acceptable predefined level of performance. Therefore, the hypotheses are inverted for the LC-CUSUM: with H_0 the process is out of control and with H_1 the process is in control.

In terms of graphical representation, the process is assumed to be unacceptable as long as the LC-CUSUM score remains above (or below if successes are indicated by an ascending graph) the limit h; the process is considered to be acceptable (the trainee has learned the procedure) when the LC-CUSUM score crosses this limit. The LC-CUSUM incorporates a holding barrier at zero that cannot be crossed. Thus, if the trainee accumulates numerous successive failures, the LC-CUSUM score will remain at zero and will not deviate far away from the limit h. Therefore, the LC-CUSUM remains responsive at all times and if, for instance, the poor performance resulted from poor technique, with improvement in technique the trainee will not have to compensate unnecessarily for all the accumulated failures and may be able to show adequate performance in due course.

In practice, the unacceptable failure rate (P_0), the acceptable failure rate (P_1; the level of performance required to indicate that the trainee is proficient) and the properties of the test have to be set. The control limit h depends on P_1 and on the properties of the test, that is the speed with which it signals that the required level of performance has been reached. When the LC-CUSUM score crosses this boundary, sufficient evidence has accumulated to indicate that the procedure has been learned. CUSUM test performances are usually expressed in terms of average run length (ARL), defined as the average number of procedures before a signal occurs under the null (ARL_0) and alternative (ARL_1) hypotheses. Detailed explanations with regard to hypotheses, formulation of LC-CUSUM, performance of the test and its limits are given in the *Appendix* (published online at www.bjs.co.uk).

The following parameters were applied to this series: for the learning curve, the hypothesis H_0 was set with $P_0 = 0{\cdot}175$ (failure rate 17·5 per cent; process out of control) and H_1 with $P_1 = 0{\cdot}1$ (failure rate 10 per cent; process in control). Computer simulations to obtain the ARL_0 and ARL_1 with different limits are displayed *Table 1* of the *Appendix*. A control limit of $h = 1{\cdot}25$ was chosen to yield ARL_0 and ARL_1 values of 74 and 32 respectively. For the standard CUSUM test, the target failure rate was set at

$P_0 = 0{\cdot}1$ (H_0; process in control) and a failure rate above $P_1 = 0{\cdot}25$ (H_1; process out of control) was considered as suboptimal performance. A control limit of $h = 2{\cdot}25$ was chosen to yield ARL_0 and ARL_1 values of 119 and 24 respectively.

Results

Overall, 532 ERCPs were performed on 237 women and 295 men with a mean age of 61 (range 17–97) years. Data on the results of cannulation were missing for three (0·6 per cent) of the 532 ERCP procedures. Fifty of the 529 ERCPs were considered unsuccessful; selective cannulation was achieved in 479 patients, giving an overall successful cannulation rate of 90·5 per cent. More details may be found in the original report[9].

The cumulative sums of failures are shown in *Fig. 1*. The success rate increased from 82·0 per cent for the first 100 procedures to 88·0, 90·0 and 95·0 per cent respectively for subsequent groups of 100 procedures. There were five failures for the last 129 cannulations, giving a success rate of 96·1 per cent.

The LC-CUSUM signalled that sufficient evidence had accumulated to indicate that the endoscopist was competent at the 79th procedure (*Fig. 2*). Until this point, 15 cannulations had failed (19 per cent), which is above the objective of a 10 per cent failure rate; however, because 16 cannulations in a row were successful, from the 64th to the 79th procedures, the LC-CUSUM signalled. The LC-CUSUM responded in a timely fashion to an improvement in the performance of the endoscopist.

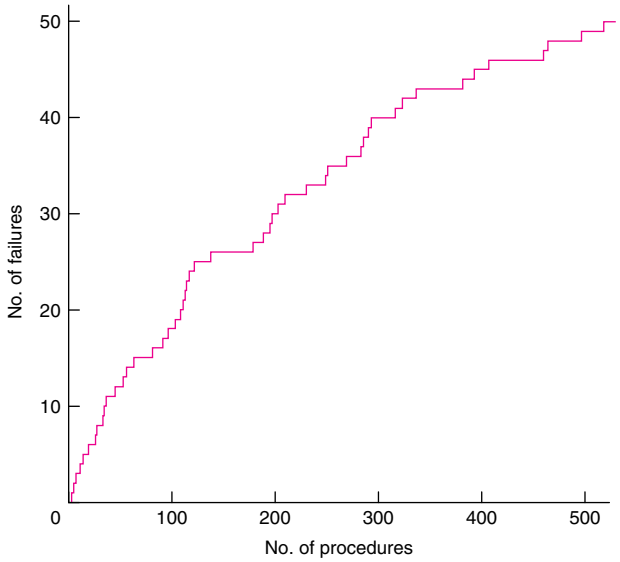

Fig. 1 Cumulative number of failures

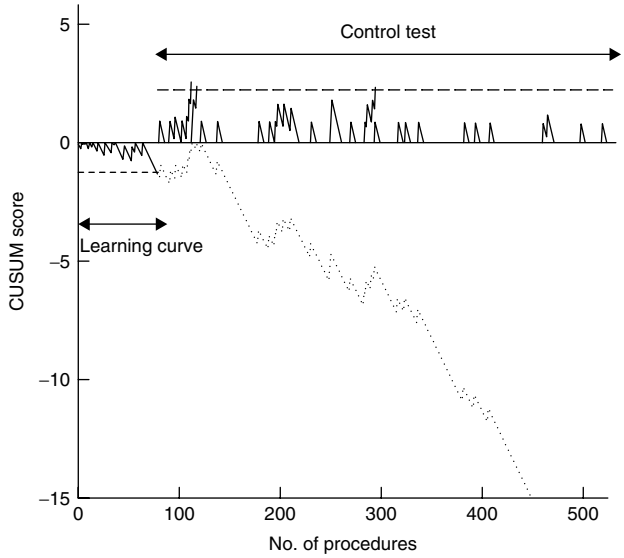

Fig. 2 Cumulative summation test for learning curve (LC-CUSUM) and standard cumulative summation (CUSUM) test results. The LC-CUSUM is applied until acceptable performance has been reached (at the 79th procedure). The standard CUSUM test is applied thereafter (control test). The dashed lines represent the limits h at $-1{\cdot}25$ for the LC-CUSUM and at $+2{\cdot}25$ for the standard CUSUM test. The dotted line represents the LC-CUSUM as if it had been continued after the 79th procedure

A standard CUSUM test was started after the 79th procedure to ensure that the operator did not deviate from optimal performance. Alarms were raised after procedure numbers 112, 117 and 292. No further alarms were raised before this series of observations ended after the 529th procedure.

Discussion

The Institute of Medicine has encouraged healthcare organizations to develop a culture of safety and create systems for continuous monitoring of patient safety[16]. As learning a new procedure carries the risk of an unacceptable standard, supervision of trainees should be mandatory until an acceptable level of performance has been reached. Deciding when a trainee has achieved this is complex as the procedure[17–19], trainee[10], tutor[20], settings[21] and required level of performance all influence the time and the number of procedures required to complete the learning process.

LC-CUSUM has been developed to enable quantitative assessment of individual performance.

The number of procedures thought necessary to learn a particular procedure varies widely. In a systematic review, Dagash and colleagues[7] found that the reported number of procedures needed to reach proficiency varied from eight to 200 for cholecystectomy, 20 to 60 for fundoplication, and 13 to 70 for colectomy. Some trainees will never be skilled enough to perform some procedures[10]. These findings highlight the need for assessment of individual competence. LC-CUSUM signals when a predefined level of performance has been achieved, regardless of the number of procedures performed.

Direct observation by a tutor and graphical representation of the learning curve are two common ways of judging individual competence, but both are likely to be subjective[10,11]. Graphical representation of a trainee's results has proved helpful in analysing the learning curve[6,9,10], with evidence of proficiency appearing as a plateau or a certain slope, depending on how the graph is constructed. This approach, however, lacks a formal test or threshold that provides objective evidence of competence. In the present study, the cumulative sums of failures were plotted and the learning curve was expected to decrease gradually to a certain slope (here, with a 10 per cent failure rate targeted, the expected slope was one in ten). Nonetheless, it is difficult to be sure exactly when an acceptable success rate was achieved. LC-CUSUM provides a good indication of when the trainee has reached the required level of performance.

Although the standard CUSUM test has been used to assess the learning curve, it is not really suitable for this purpose[15]. During the learning process, the upper boundary limit of this test is often crossed, sometimes two or three times, so setting a limit makes little sense[12,22]. Moreover, once the limit has been crossed, the statistical properties of the test are lost and the hypothesis of interest cannot be tested. Other methods, such as the exponentially weighted moving average[23], the risk-adjusted CUSUM test[24] and the cumulative risk-adjusted mortality chart[25], have the same limitations. LC-CUSUM has been designed specifically to overcome these problems. The different parameters (target value, control limits and ARL) allow the test to be adapted to any procedure and any level of performance considered necessary. Once a trainee shows evidence of satisfactory performance, he or she may be allowed to perform the procedure without supervision. Another advantage of such monitoring is that the trainee will not lose the accumulated evidence of performance when moving from one department to another, as is often the case when the assessment is based on observations by a tutor. Monitoring of trainees with LC-CUSUM may be done by continuous implementation of a personal log book. Supervision during a procedure should not be stopped unless adequate performance has been demonstrated; thereafter monitoring may be continued with a standard CUSUM test, and supervision recommended only when the presence of additional difficulties, such as co-morbidities, makes it necessary.

The present study has several limitations. The acceptable and unacceptable failure rates and the performance of the test were chosen arbitrarily. These limits must be set rigorously, and the risks balanced under the null and alternative hypotheses. Nonetheless, these values can be modified according to expert recommendations. In the original study by Schlup and colleagues[9], had the LC-CUSUM been continued after the 79th procedure it would have gone over the lower boundary line for a few more procedures indicating that performance had not reached the required level. With a more conservative limit, it would have signalled later during monitoring. Once a trainee has reached an acceptable level of performance, ideally self-monitoring should be continued with a standard CUSUM test to ensure that the level of performance is maintained. This issue has been referred to previously as the 'retention curve'[26].

Monitoring rare events such as death after appendicectomy or failure to complete a cholecystectomy may show poor statistical properties owing to the expected limited number of failures. However, one may choose to define a list of additional events ('near misses') in addition to the event of interest ('miss') as a proxy for inadequate performance. This approach has been implemented successfully in cardiothoracic surgery to monitor the outcome of complex cardiac procedures[3].

Implementation of such monitoring requires time and resources, but the cost seems justified in the light of the increasing requirement to assess performance. LC-CUSUM has been used to assess the performance of a single individual in the present study, but it may prove useful for monitoring the introduction of a new procedure, although it would be valuable in such a setting only if corrective actions could be implemented. It may also prove useful to expert societies responsible for developing guidelines for good practice.

References

1 The Inquiry into the management of care of children receiving complex heart surgery at the Bristol Royal Infirmary. Final report. http://www.bristol-inquiry.org.uk/ [accessed 17 March 2006].

2 Institute of Medicine. *Crossing the Quality Chasm: a New Health System for the 21st Century*. National Academy Press: Washington, DC, 2001.
3 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. *J Thorac Cardiovasc Surg* 1994; **107**: 914–923.
4 Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *Int J Qual Health Care* 2003; **15**: 7–13.
5 Nizard RS, Porcher R, Ravaud P, Vangaver E, Hannouche D, Bizot P *et al*. Use of the Cusum technique for evaluation of a CT-based navigation system for total knee replacement. *Clin Orthop Relat Res* 2004; **425**: 180–188.
6 Young A, Miller JP, Azarow K. Establishing learning curves for surgical residents using Cumulative Summation (CUSUM) Analysis. *Curr Surg* 2005; **62**: 330–334.
7 Dagash H, Chowdhury M, Pierro A. When can I be proficient in laparoscopic surgery? A systematic review of the evidence. *J Pediatr Surg* 2003; **38**: 720–724.
8 Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. *Ann Surg* 1995; **222**: 735–742.
9 Schlup MM, Williams SM, Barbezat GO. ERCP: a review of technical competency and workload in a small unit. *Gastrointest Endosc* 1997; **46**: 48–52.
10 Van Rij AM, McDonald JR, Pettigrew RA, Putterill MJ, Reddy CK, Wright JJ. Cusum as an aid to early assessment of the surgical trainee. *Br J Surg* 1995; **82**: 1500–1503.
11 Elliot DL, Hickam DH. Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. *JAMA* 1987; **258**: 3405–3408.
12 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.
13 Van Dobben de Bruyn CS. *Cumulative Sum Tests: Theory and Practice*. Griffin's Statistical Monographs and Courses no. 24. Hafner: London, 1968.
14 Williams SM, Parry BR, Schlup MM. Quality control: an application of the cusum. *BMJ* 1992; **304**: 1359–1361.
15 Biau DJ, Resche-Rigon M, Godiris-Petit G, Nizard RS, Porcher R. Quality control of surgical and interventional procedures: a review of the CUSUM. *Qual Saf Health Care* 2007; **16**: 203–207.
16 Kohn LT, Corrigan J, Donaldson MS (eds). *To Err Is Human: Building a Safer Health System*. National Academy Press: Washington, DC, 2000.
17 Archibeck MJ, White RE Jr. Learning curve for the two-incision total hip replacement. *Clin Orthop Relat Res* 2004; **429**: 232–238.
18 Schauer P, Ikramuddin S, Hamad G, Gourash W. The learning curve for laparoscopic Roux-en-Y gastric bypass is 100 cases. *Surg Endosc* 2003; **17**: 212–215.
19 Schlachta CM, Mamazza J, Seshadri PA, Cadeddu M, Gregoire R, Poulin EC. Defining a learning curve for laparoscopic colorectal resections. *Dis Colon Rectum* 2001; **44**: 217–222.
20 Ahlberg G, Kruuna O, Leijonmarck CE, Ovaska J, Rosseland A, Sandbu R *et al*. Is the learning curve for laparoscopic fundoplication determined by the teacher or the pupil? *Am J Surg* 2005; **189**: 184–189.
21 Reichenbach DJ, Tackett AD, Harris J, Camacho D, Graviss EA, Dewan B *et al*. Laparoscopic colon resection early in the learning curve: what is the appropriate setting? *Ann Surg* 2006; **243**: 730–735.
22 Kestin IG. A statistical approach to measuring the competence of anaesthetic trainees at practical procedures. *Br J Anaesth* 1995; **75**: 805–809.
23 Shehab RL, Schlegel RE. Applying quality control charts to the analysis of single-subject data sequences. *Hum Factors* 2000; **42**: 604–616.
24 Tekkis PP, Fazio VW, Lavery IC, Remzi FH, Senagore AJ, Wu JS *et al*. Evaluation of the learning curve in ileal pouch–anal anastomosis surgery. *Ann Surg* 2005; **241**: 262–268.
25 Sismanidis C, Bland M, Poloniecki J. Properties of the cumulative risk-adjusted mortality (CRAM) chart, including the number of deaths before a doubling of the death rate is detected. *Med Decis Making* 2003; **23**: 242–251.
26 Matz R. The learning curve. *JAMA* 1994; **271**: 825.