International Journal for Quality in Health Care 2013; **Volume 25, Number 3:** pp. 322–330
Advance Access Publication: 4 April 2013
10.1093/intqhc/mzt028

Assessment of competency in clinical measurement: comparison of two forms of sequential test and sensitivity of test error rates to parameter choice

ANDREW J. SIMS[1,3], KIM KELTIE[1,3], JULIE BURN[1] AND STEPHEN C. ROBSON[2,3]

[1]Department of Medical Physics, Newcastle upon Tyne Hospitals NHS Foundation Trust, UK, [2]Department of Fetal Medicine, Newcastle upon Tyne Hospitals NHS Foundation Trust, UK and [3]Institute of Cellular Medicine, Newcastle University, UK

Address reprint requests to: Andrew J. Sims, Regional Medical Physics Department, Newcastle upon Tyne Hospitals NHS Foundation Trust, Freeman Hospital, Newcastle upon Tyne NE7 7DN, UK. Tel: +44-191-2448738; Email: andrew.sims@nuth.nhs.uk

Accepted for publication 20 February 2013

Abstract

Objective. To assess clinical measurement competency by two sequential test formulations [resetting sequential probability ratio test (R-SPRT) and learning curve cumulative summation (LC-CUSUM)].

Design. Numerical simulation and retrospective observational study.

Setting. Obstetric ultrasound department.

Participants. Cohorts of 10 000 simulated trainees and 62 obstetric sonographers training in nuchal translucency (NT) measurement at the 11–14-week pregnancy scan with limited case availability.

Intervention. Application of LC-CUSUM and R-SPRT to clinical measurement training.

Main Outcome Measures. Proportions of real trainees achieving competency by LC-CUSUM and R-SPRT, proportions of simulated competent trainees not achieving competency (Type I error), proportions of simulated incompetent trainees achieving competency (Type II error), distribution of case number required to achieve competency (run length) and frequency of resets.

Results. For simulated cohorts, significant differences in run-length distribution and true test error rates were found between the R-SPRT and LC-CUSUM tests with equivalent parameters. Increasing the cases available to each trainee reduced the Type I error rate but increased the Type II error rate for both sequential tests for all choices of unacceptable failure rate. Discontinuities in the proportion of trainees expected to be test competent were found at critical values of unacceptable failure rate.

Conclusions. With equivalent parameters, the R-SPRT and LC-CUSUM formulations of sequential tests produced different outcomes, demonstrating that the choice of test method, as well as the choice of parameters, is important in designing a training scheme. The R-SPRT detects incompetence as well as competence and may indicate need for further training. Simulations are valuable in estimating the proportions of trainees expected to be assessed as competent.

Keywords: quality control, training, computer simulation, outcome assessment, statistical

Introduction

The benefit of sequential testing to assess training outcomes is to provide a form of statistical quality control that can quickly identify competence or incompetence, and to provide a graphical method of monitoring performance over time without the need for repeated hypothesis testing or predetermined sample number [1]. Sequential statistical tests have been applied when assessing the competency of supervised trainees learning surgical procedures including ERCP [2] and coronary bypass graft [3] and in making ultrasound measurements including foetal weight [4] based on foetal biometry estimates [5, 6]. These tests considered the outcome of each case undertaken by each trainee to be dichotomous, with each success assigned a specific score and each failure an alternative score [7]. Trainees were considered competent

when their cumulative score, plotted against case number, crossed a pre-defined barrier [8]. Different forms of sequential test have been used to define the scores associated with success and failure, the cumulative score and the position of the barriers which mark competence and incompetence.

Two examples of such sequential tests are the resetting sequential probability ratio test (R-SPRT), which is a sequence of SPRTs [8], and the learning curve cumulative summation (LC-CUSUM) test [2], both described in the Appendix. There is little consensus in the literature on which variant is the most appropriate for particular applications [1, 8].

Parameters common to both tests are the acceptable failure rate (p_0), at which a competent practitioner is expected to make errors, and an error rate above which a practitioner would be considered incompetent, the unacceptable failure rate (p_1) [9].

Two additional parameters are required to define an R-SPRT: the Type I error parameter (α^*), which defines the proportion of competent trainees expected to be incorrectly classified as incompetent in any single SPRT, and the Type II error parameter (β^*), which defines the proportion of incompetent trainees expected to be incorrectly classified as competent in any single SPRT [8]. Because an R-SPRT is a sequence of SPRTs, the true Type I and Type II error rates of the test, α and β, normally differ from the expected values, α^* and β^*, respectively [8]. Most clinical training applications of R-SPRT have set both α^* and β^* equal to 10%. The four initial parameters (p_0, p_1, α^* and β^*) of an R-SPRT define two barriers; a terminating barrier (h_0) which, when crossed, indicates that competency has been achieved, and a resetting barrier (h_1), which when crossed, indicates incompetence, and the cumulative score is reset to zero and the test continues (i.e. another SPRT begins) [10].

In an LC-CUSUM, the height of the resetting barrier is fixed at zero [8]. When the cumulative score crosses this barrier, it is reset to zero and the test continues. In contrast to the R-SPRT, crossing this barrier does not necessarily indicate incompetence (e.g. a trainee's score would cross this barrier following an early failure). In common with the R-SPRT, the LC-CUSUM has a terminating barrier, which, when crossed, indicates competency. However in an LC-CUSUM, the position of this barrier is a free parameter chosen to suit the needs of the test. Biau *et al.* [2] recommended using simulation of average run lengths (ARLs) to inform the choice of barrier height. Examples of an R-SPRT and LC-CUSUM are shown in Fig. 1.

An additional variable that influences the outcome of both R-SPRT and LC-CUSUM tests is the practical limit in number of cases available to each trainee (N_{max}). Both the R-SPRT and LC-CUSUM tests terminate only when the terminating barrier is crossed. Thus, when N_{max} is finite, there are two possible outcomes for each trainee: 'competent' (i.e. cumulative score crossed the terminating barrier) or 'not yet competent' (i.e. cumulative score did not cross the terminating barrier within N_{max} cases). If provided with a large enough number of cases (i.e. $N_{max} \to \infty$), all trainees would eventually be considered competent by chance. In consequence, as N_{max} increases, the true Type I error rate, α,

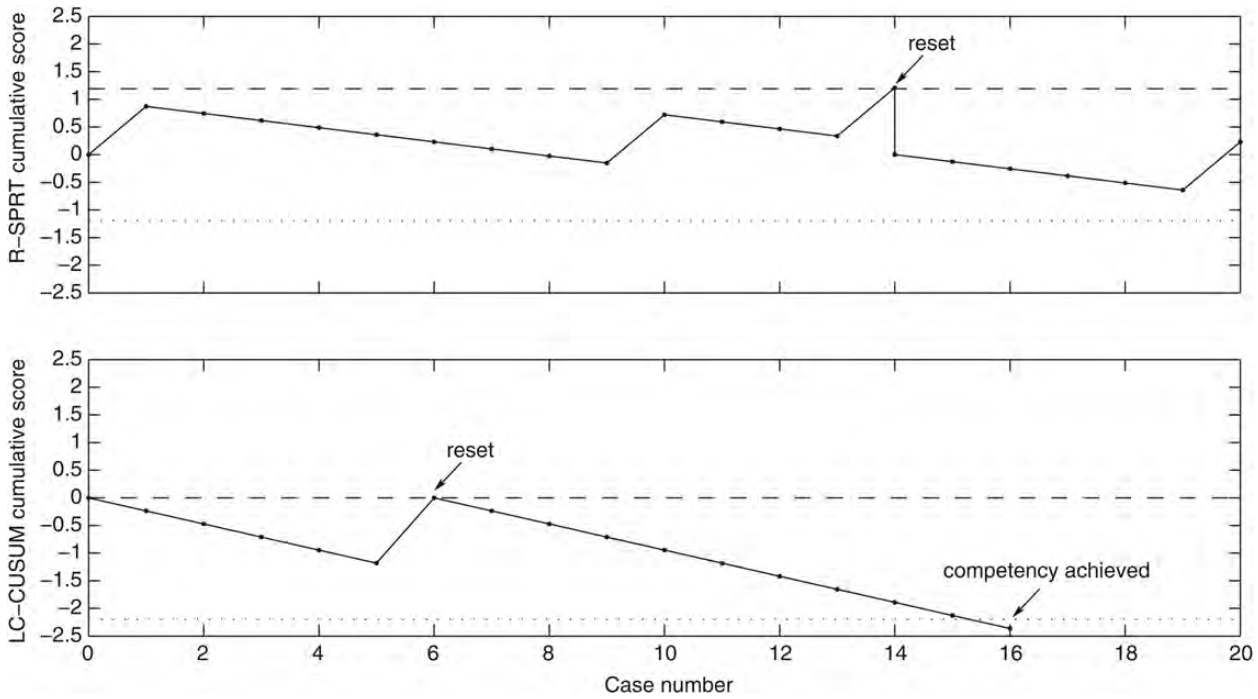

Figure 1 *Upper panel*: example of an R-SPRT for a trainee 'not yet competent' after 20 cases. *Lower panel*: example of an LC-CUSUM for a different 'competent' trainee after 16 cases. Dashed lines are resetting barriers; dotted lines are terminating barriers.

tends to zero (i.e. fewer truly competent trainees will fail to achieve test competence) and the true Type II error rate, β, tends to one (i.e. more truly incompetent trainees will eventually achieve test competence) [8].

The values chosen for the initial parameters of an R-SPRT (p_0, p_1, α^* and β^*) or LC-CUSUM (p_0, p_1, h_0), together with the case limit N_{max}, define the ability of these sequential tests to distinguish correctly between trainees in a cohort who are competent from those who are not yet competent. However, previous studies [2–6, 11] did not calculate 'a priori' the efficacy of their particular test design to justify their choices of parameter and were unable to compare observed with expected performance.

The objectives of this study were: (i) to directly compare run-length distributions of R-SPRT with LC-CUSUM tests using simulation; (ii) to assess the effect of varying the test parameters p_1 and N_{max} on the true Type I and Type II error rates of R-SPRT and LC-CUSUM tests; (iii) to use simulation to calculate expected training outcomes, assessed by R-SPRT and LC–CUSUM, of a cohort of trainees learning to ultrasonically measure nuchal translucency (NT, foetal neck measurement offered to all pregnant women at the 11–14-week scan to screen for Down's syndrome) and (iv) to compare simulated outcomes with actual training outcomes.

Methods

Simulated cohorts of trainees

The methodology of the R-SPRT and LC-CUSUM techniques, described in the Appendix, were applied to simulated cohorts of 10 000 trainees, using Matlab (Mathworks, Ltd, Cambridge, UK) [12]. Each trainee was assumed to have a probability, r, of making an error in a single case. The outcome of each individual case was modelled as a Bernoulli random variable with success probability $1-r$ [13]. The cumulative score, S, for each simulated trainee, was calculated for each case in the series, until it crossed the terminating barrier (h_0), or until the maximum number of cases, N_{max}, was reached. All simulations were repeated 100 times in order to reduce statistical variation, and the mean number of trainees achieving competency was recorded for each combination of initial parameters.

Cohort of trainees learning to measure NT

Between August 2009 and November 2010, 62 obstetric sonographers were enrolled in a formal programme of training in the measurement of NT at the Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. Each trainee spent 1 week in the obstetric ultrasound department and measured NT in a series of pregnant women attending for an 11–14-week scan. Supervision of trainees was provided by two senior obstetric sonographers with >5 years' experience of measuring NT, both competent in measuring NT (as defined by accreditation by the Fetal Medicine Foundation [14]). For each case, both the trainee and trainer measured NT three times and the measurement was considered successful if the difference between mean trainee and mean trainer measurements was less than a defined value. R-SPRT and LC-CUSUM tests were applied retrospectively to the cumulative score, S, after each trainee had completed their placement. The retrospective analysis of NT measurements did not influence the management of the patient, or the training outcome of any trainee.

Training outcomes

For each cohort, the following outcomes were recorded: the proportion of trainees reaching competency (i.e. S crossed the terminating barrier); the proportion of trainees not yet reaching competency (i.e. S did not cross the terminating barrier within N_{max} cases); the distribution of the number of cases required for trainees to achieve competency (i.e. run-length distribution); the distribution of the number of times S crossed the resetting barrier (of practical value for R-SPRT only).

Choice of test parameters

In this study, which is concerned with the measurement of continuous clinical variables, it was assumed that a trainee measurement was a success if it agreed with that of a trained observer within a tolerance limit, and a failure otherwise. The choice of tolerance and the value of acceptable failure rate were based on published inter-observer agreement between trained observers. For foetal NT measurements, Pandya *et al.* [15] reported that 95% of mean measurements by pairs of trained observers agreed within 0.44 mm. Therefore, for both R-SPRT and LC-CUSUM tests, each case was ascribed a 'success' when a trainee measured within 0.44 mm of that by a trained observer, and a 'failure' otherwise, and the acceptable failure rate, p_0, was set to 5%.

For simulations where the value of unacceptable failure rate, p_1, was required to be fixed, its value was chosen to give the same p_1/p_0 ratio applied in previous studies that also assessed trainee competency using the R-SPRT technique [9].

For the R-SPRT, the Type I (α^*) and Type II (β^*) initial parameters were both set to 10%. This simplifies the graphical presentation of the R-SPRT plot by placing the terminating and resetting barriers at equal and opposite distances from the starting position, enabling the plot to be examined in a continuous fashion, while also achieving a compromise between the more conventional choices of 5 and 20%, respectively [9, 11].

For LC-CUSUM tests, the position of the terminating barrier, h_0, was chosen to make the minimum number of cases required to achieve competency (minimum run length) equivalent to that of an R-SPRT test with the same initial parameters (α^*, β^*, p_0 and p_1).

For the simulated cohort, three values of probability of error, r, were used. These were set to be equal to: p_0 (a competent cohort), p_1 (an incompetent cohort) and the threshold probability ($r = s$) at which the mean increment in cumulative score, S, across many cases is zero (Appendix).

Comparison of run-length distributions between R-SPRT and LC-CUSUM tests

Run-length distributions for R-SPRT and LC-CUSUM tests, with $p_1 = 25\%$ [4] and unlimited cases per trainee ($N_{max} = \infty$), were compared using simulated cohorts of competent ($r = p_0$) and incompetent ($r = p_1$) trainees using the χ^2 goodness of fit test [16]. With these parameters ($\alpha^* = 0.1$, $\beta^* = 0.1$, $p_0 = 0.05$, $p_1 = 0.25$), the minimum run length to achieve test competence was 10 cases for both tests (Appendix).

Variation of true Type I and Type II error rates with p_1 and N_{max}

From the simulated cohort, the proportion of competent trainees ($r = p_0$) not achieving test competency was used to estimate the true Type I error rate (α) and the proportion of incompetent trainees ($r = p_1$) achieving test competency was used to estimate the true Type II error rate (β). These true error rates (α, β) were calculated for LC-CUSUM and R-SPRT for different values of p_1 (10, 15, 20, 25 and 30%) and N_{max} (range 10–3000) using simulated cohorts.

Comparison of observed and expected test outcomes

Expected outcomes were calculated for R-SPRT and LC-CUSUM tests, for simulated cohorts of trainees each provided with a limited number of cases ($N_{max} = 25$, to allow comparison with the ultrasonographic trainees). Simulations were run for three single case failure probabilities ($r = p_0$, s and p_1) and values of unacceptable failure rate ranging between 15 and 30%. For the R-SPRT cases at $p_1 = 25\%$, the number of crossings of the resetting barrier was recorded for each trainee to give an expected distribution of resets. This distribution was scaled to the number of NT trainees to give the expected distribution of resets for comparison with that cohort.

Observed outcomes for the ultrasonographic trainees were calculated by retrospectively applying the R-SPRT and LC-CUSUM tests to trainee and trainer NT measurements, for values of unacceptable failure rate ranging between 15 and 30%. For R-SPRT calculations, the observed distribution of resets was also calculated for a fixed unacceptable failure rate, $p_1 = 25\%$.

Results

Comparison of run-length distributions between R-SPRT and LC-CUSUM tests

The ARL for competent trainees ($r = p_0$) was 15.3 cases [median 10; inter-quartile range (IQR) 10–18] for R-SPRT and 13.2 cases (median 10; IQR 10–16) for LC-CUSUM. For incompetent trainees with $r = p_1$, the ARL was 114.9 cases (median 82; IQR 37–157) for R-SPRT and 66.7 cases (median 49; IQR 24–89) for LC-CUSUM (Fig. 2). The distributions of run lengths described by R-SPRT and LC-CUSUM were found to be significantly different ($P < 0.01$).

Variation of true Type I and Type II error rates with p_1 and N_{max}

Figure 3 shows the true Type I and Type II error rates for ranges of p_1 and N_{max}, obtained from simulation, for $p_0 = 5\%$, $\alpha^* = 10\%$ and $\beta^* = 10\%$. Increased N_{max} was found to decrease the true Type I error and increase the true Type II error for both R-SPRT and LC-CUSUM tests.

For the same example shown by the dotted lines and shading in Fig. 2, with a case limit of 25 and unacceptable failure rate of 25%, the R-SPRT had a true Type I error rate of 17.0% and a Type II error rate of 17.6% (β). For the same test parameters, the LC-CUSUM test had a true Type I error rate of 3.7% and a Type II error rate of 27.2%.

Comparison of observed and expected test outcomes

Application of both R-SPRT and LC-CUSUM tests, with the same values of p_0, p_1 and minimum run length, to the cohort of NT trainees resulted in a higher proportion of trainees achieving competency with the LC-CUSUM tests compared with R-SPRTs (Fig. 4). This was observed for all unacceptable failure rates. The standard deviation in the number of simulated trainees reaching competency between the 100 separate simulations was <0.5%. The median number of cases measured by the NT training cohort was 25 (range 15–31), allowing direct comparison of proportions of successful trainees with the simulated cohorts ($N_{max} = 25$). By both types of test, the proportions of NT trainees achieving test competence were consistent with a trainee error probability of less than the threshold probability, s.

Near some critical values of p_1, relatively small increases in unacceptable failure rate resulted in stepped increases in the proportions of simulated trainees achieving competency by both LC-CUSUM and R-SPRT tests, Fig. 4. These critical values of p_1 occurred at points where the minimum run length, m, changed by one case, with larger discontinuities observed when m was close to the most frequent run lengths. These critical values of p_1 are calculated by the relation $(1 - p_1) = (1 - p_0)\ ((1-\alpha^*)/\beta^*)^{-1/m}$, Fig. 4.

The distributions of resets for the R-SPRT method for simulated (normalized to 62 trainees) and NT training cohorts are shown in Fig. 5; for example, 20 of the 62 NT trainees reached the resetting barrier once. The proportions of NT trainees with 0, 1, 2, 3 and 4 resets were consistent with a trainee error probability less than the threshold probability of error, s.

Discussion

Previous studies have demonstrated the potential for using sequential tests to assess training in surgical procedures and

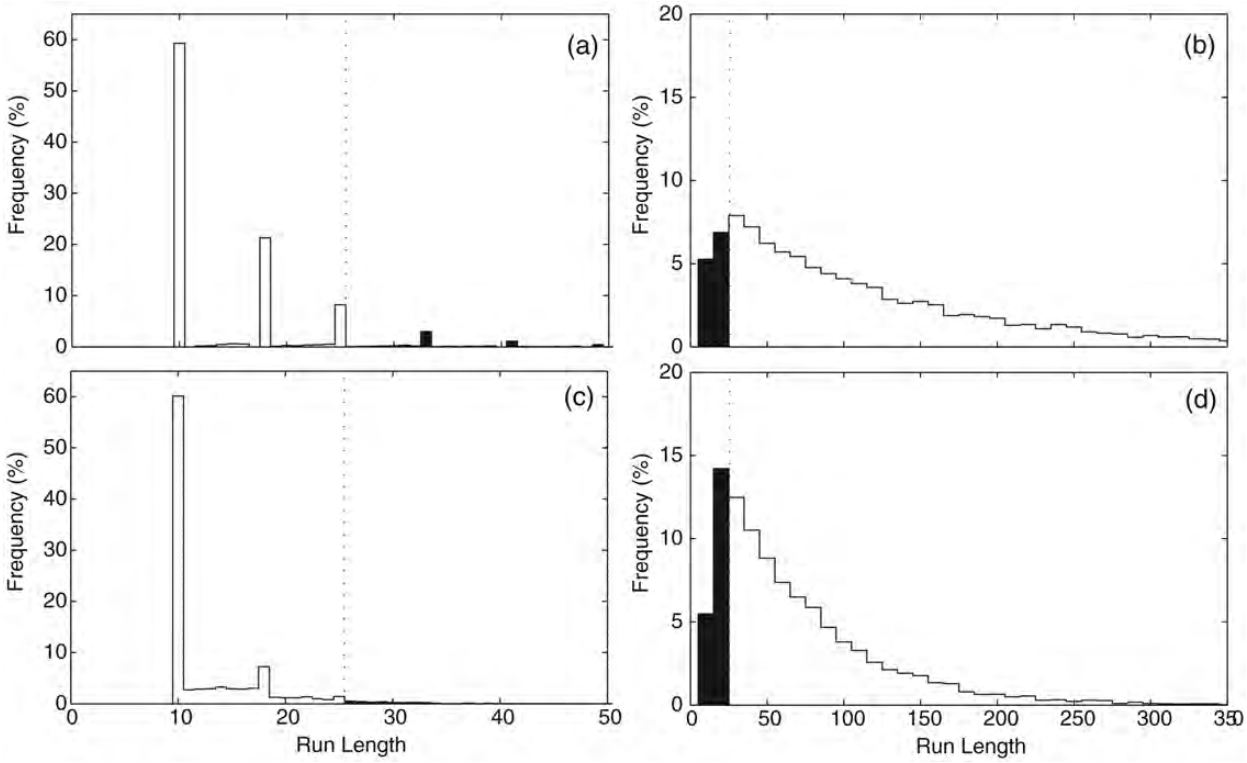

Figure 2 Distributions of number of cases required to reach competency (run-length distributions) for simulated cohorts of 10 000 trainees with $p_0 = 5\%$, $p_1 = 25\%$ and no limit to the number of cases available to each trainee. *Upper panels (a, b)*: R-SPRT. *Lower panels (c, d)*: LC-CUSUM. *Left panels (a, c)*: competent trainees make errors in 5% of cases, i.e. $r = p_0$. *Right panels (b, d)*: incompetent trainees make errors in 25% of cases, i.e. $r = 0.25$. The dotted lines illustrate the effect of limiting each trainee to 25 cases. The shaded bars are trainees classified in error, after limiting to 25 cases, with the shaded proportions giving true Type I errors of 17.0% (*a*) and 3.7% (*c*) and true Type II errors of 17.6% (*b*) and 27.2% (*d*).

clinical measurement [1–5, 10]. However, in this study, we have shown that simulation can be used to estimate expected performance and have compared it with the outcome from an observed cohort of trainees learning to measure NT as part of an 11–14-week pregnancy scan.

We found that, for equivalent test parameters, R-SPRT and LC-CUSUM tests had different outcomes. They had a different run-length distribution, which led to different behaviour when the limit to the number of cases available to each trainee was finite. The LC-CUSUM test yielded a smaller true Type I error rate and a larger true Type II error rate than an equivalent R-SPRT with the same case number limit.

The LC-CUSUM test is similar to an R-SPRT with a resetting barrier at zero; however, there are fundamental differences in their application [8]. By choosing parameters that gave equal minimum run length for both types of test, we ensured that the terminating barriers were equivalent and the differences we observed between the types of test were due to the difference in position of the resetting barrier. However, on average, trainees crossed the terminating barrier (i.e. achieve competence) more quickly with an LC-CUSUM test than with the equivalent R-SPRT. This is a consequence of trainees being penalized for errors early on in training with the R-SPRT test, where a string of consecutive successes is required to compensate for previous mistakes. This is not observed with the LC-CUSUM test, where errors early on in training reset the score to zero, effectively reinitiating the training process, and thus, no additional successes are required in order to reach competency.

We have calculated the true Type I and Type II error rates associated with R-SPRT and LC-CUSUM tests with an acceptable failure rate of 5%, for a range of unacceptable failure rates and for various limits to the number of cases available to each trainee. These error rates relate to the expected number of false positives (i.e. incompetent trainees who pass the test) and false negatives (i.e. number of competent trainees who fail the test) and can be used to inform the choice of parameters when designing a programme for training in clinical measurement, in which formal assessment is by sequential test.

For both sequential tests, the true Type I error rate, α, tends to zero and Type II error rate, β, tends to one when $N_{max} >> ARL$ for the simulated cohort. However, for smaller values of N_{max}, an LC-CUSUM test yielded smaller α and larger β error rates than the equivalent R-SPRT. When applying a sequential test to a training scheme, both Type I and Type II error rates should be minimized. Increasing the

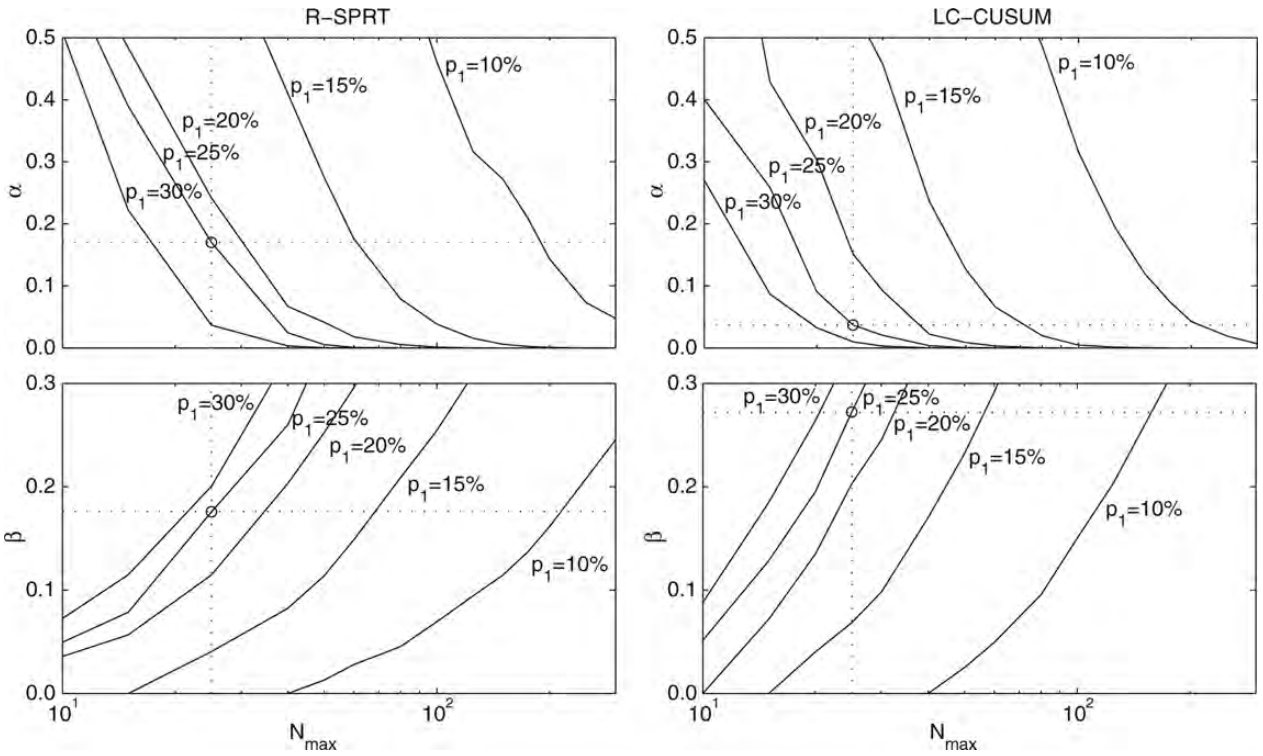

Figure 3 The true Type I (α) and Type II (β) error rates associated with R-SPRT (left hand panels) and LC-CUSUM (right hand panels) tests for using initial parameter settings $p_0 = 5\%$, $\alpha^* = 10\%$, $\beta^* = 10\%$ when unacceptable failure rate, p_1, and case limit, N_{max} are varied. The dotted lines indicate measured test error rates for $p_1 = 25\%$ with each trainee limited to 25 cases.

Type I error rate causes fewer truly competent trainees to reach competency, and thus, a longer training period would be required, at a consequence of increased training cost. Increasing the Type II error rate would result in a larger number of truly incompetent trainees being deemed competent, potentially enabling incompetent trainees to carry out subsequent inaccurate clinical measurements in clinical practice without supervision.

The sensitivity of R-SPRT and LC-CUSUM tests to small changes in the choice of unacceptable failure rate, p_1, was a finding revealed by our use of simulation and one that has not, to our knowledge, been reported previously. The presence of discontinuities at critical values of p_1 was associated with step changes in minimum run length. Figure 2 illustrates why this is so; the distribution of run lengths (i.e. number of cases needed to become competent) is not smooth when the probability of error lies near p_0, with run lengths that are frequent (e.g. 10 in the example for R-SPRT), impossible (e.g. 11) or unlikely (e.g. 12). If a small decrease in p_1 leads to an increase in minimum run length, the run length of other common values in the distribution will also change, e.g. the peak at $N = 25$ may move to $N = 26$. If the case limit is close to one of these common values (as in the example of Fig. 2), then there will be a large change in the proportion of trainees expected to reach competency.

In practical terms, N_{max} should be chosen to avoid it lying near a peak in the run-length distribution; equivalently, p_1 should be chosen to avoid it lying near a critical value, otherwise the consequence of some trainees failing to measure N_{max} cases will have a disproportionate effect on α and β.

For the purposes of assessing training, R-SPRT has a potential advantage over LC-CUSUM. Interpreted as a Wald test, the points at which a trainee's R-SPRT cumulative score exceeds the resetting barrier may be considered indication of incompetence. The number of resets could, therefore, be used to trigger further intervention in their training, e.g. to indicate the need for revision of theoretical principles or initiation of retraining. However, the number of resets allowed, as with the choice of sequential test methodology and parameters, is a clinical decision, based on the rigour of training wanted, and not a statistical one.

Our study had limitations. For all calculations, we used 10% for the test error parameters α^* and β^*, following previous studies; further simulation would be needed to create comprehensive tabulation of test performance for alternative values of these parameters. Trainees in the NT cohort were constrained by the number of cases available to them with a median of 25 (range 15–31). The case number limit chosen for simulation (25) was an approximation to the constraints placed on the NT cohort.

Simulations of large training cohorts can estimate run-length distributions, indicate sequential test sensitivity, inform the length of training to optimize sequential test efficacy and estimate values of the true Type I and Type II error rates for a given set of initial parameters. In utilizing a simulated cohort,

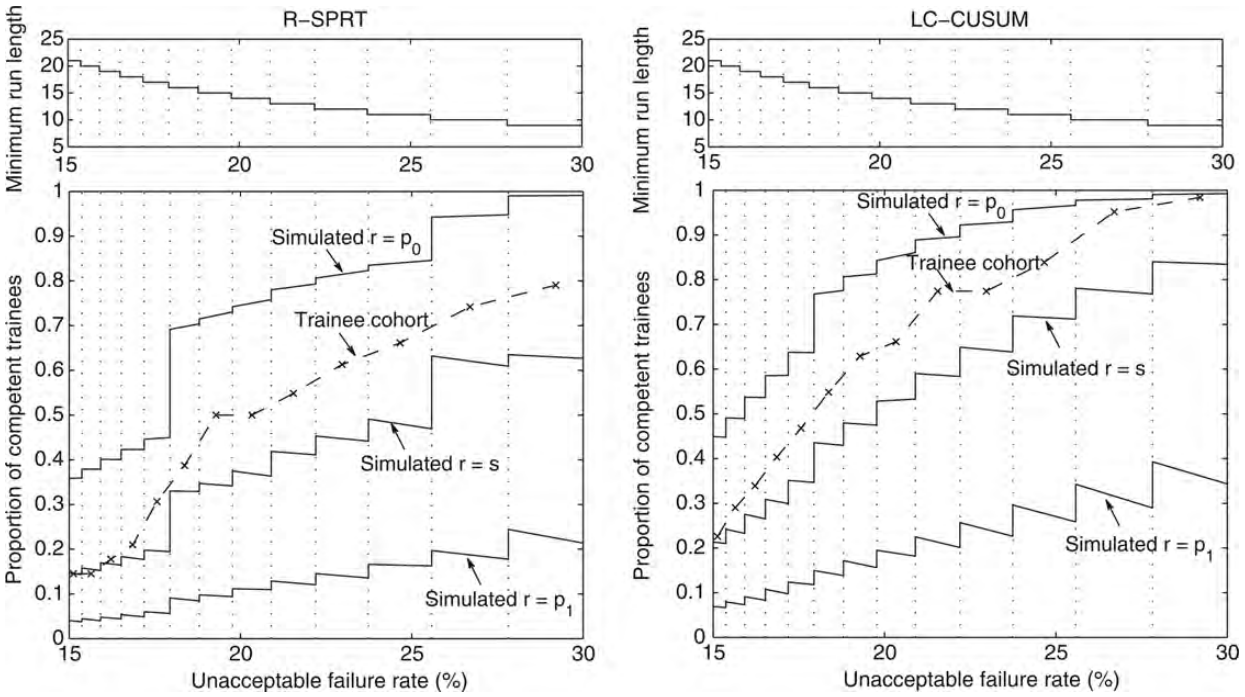

Figure 4 *Upper panels*: minimum run length needed to cross the competency barrier for a range of unacceptable failure rates (in order to fairly compare the R-SPRT and LC-CUSUM techniques, the parameters were chosen such that the minimum run lengths were the same and thus by definition the upper panels are identical). *Lower panels*: training outcomes (proportion of competent trainees) compared for a simulated cohort of 10 000 trainees (solid lines) and the NT training cohort (dashed lines), for a range of unacceptable failure rates (p_1). Outcomes for the simulated cohort are given for three single case failure probabilities (p_0, s and p_1) with each trainee limited to 25 cases. *Left panels*: R-SPRT test with parameters $p_0 = 5\%$, $\alpha^* = 10\%$, $\beta^* = 10\%$. *Right panels*: LC-CUSUM test with $p_0 = 5\%$ and h_0 chosen to give the same minimum run length as an R-SPRT with $p_0 = 5\%$, $\alpha^* = 10\%$, $\beta^* = 10\%$. Vertical lines (dots) are drawn at critical values of p_1 associated with integral changes in minimum run length.

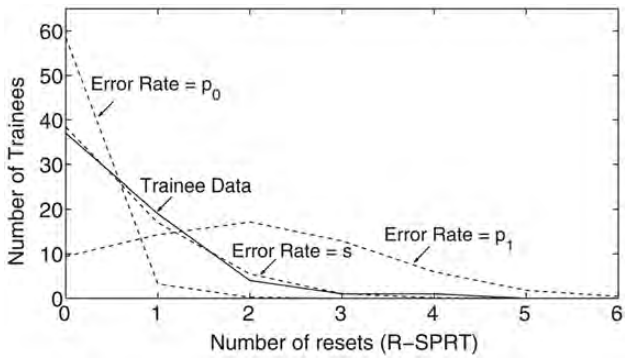

Figure 5 The normalized distribution of the number of resets for an R-SPRT test, with $p_0 = 5\%$ and $p_1 = 25\%$, for simulated cohorts of 10 000 trainees (dashed lines) with three single case failure probabilities (p_0, s and p_1) and each trainee limited to 25 cases, compared with the 62 NT trainees (solid line).

it is possible to consider all of these outcomes to inform the design and highlight intrinsic limitations of sequential testing when applied to a training scheme prospectively.

In practice, designing a programme for training in clinical measurement, including assessment by sequential testing, requires a trade-off between the practical limit to the number of cases that can be made available for each trainee and the unacceptable failure rate, at which an observer would be considered incompetent. In this study we have provided results that demonstrate the effect of varying case limit and unacceptable failure rate, and have described a technique for simulation that can be used to assess the efficacy of R-SPRT and LC-CUSUM tests with a given choice of test parameters.

Acknowledgements

We thank Angela Bradley and Vikki Smith for their help in conducting this study.

Funding

This work was supported by a National Institute of Health Research Flexibility and Sustainability award administered

by the Newcastle upon Tyne Hospitals NHS Foundation Trust.

References

1. Norris A, McCahon R. Cumulative sum (CUSUM) assessment and medical education: a square peg in a round hole. *Anaesthesia* 2011;**66**:243–54.
2. Biau DJ, Williams SM, Schlup MM *et al*. Quantitative and individualised assessment of the learning curve using LC-CUSUM. *Br J Surg* 2008;**95**:925–9.
3. Grunkemeier GL, Wu YX, Furnary AP. Cumulative sum techniques for assessing surgical results. *Ann Thorac Surg* 2003;**76**:663–7.
4. Weerasinghe S, Mirghani H, Revel A *et al*. Cumulative sum (CUSUM) analysis in the assessment of trainee competence in fetal biometry measurement. *Ultrasound Obstet Gynecol* 2006;**28**:199–203.
5. Cruz-Martinez R, Figueras F, Moreno-Alvarez O *et al*. Learning curve for lung area to head circumference ratio measurement in fetuses with congential diaphragmatic hernia. *Ultrasound Obstet Gyncol* 2010;**36**:32–6.
6. Balsyte D, Schäffer L, Burkhardt T *et al*. Continuous independent quality control for fetal ultrasound biometry provided by the cumulative summation technique. *Ultrasound Obstet Gynecol* 2010;**35**:449–55.
7. Spiegelhalter D, Grigg O, Kinsman R *et al*. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *Int J Qual Health Care* 2003;**15**:7–13.
8. Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Meth Med Res* 2003;**12**:147–70.
9. 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**:433–8.
10. Davies OL. *The Design and Analysis of Industrial Experiments*. London: Oliver and Boyd, 1954.
11. Kestin IG. Commentaries: a statistical approach to measuring the competence of anaesthetic trainees at practical procedures. *Br J Anaesth* 1995;**75**:805–809.
12. Burton A, Altman DG, Royston P *et al*. The design of simulation studies in medical statistics. *Statist Med* 2006;**25**: 4279–4292.
13. Evans M, Hastings N, Peacock B. *Statistical Distribution*. 3rd edn. New York: Wiley, 2000: 31–33.
14. Fetal Medicine Foundation. USA: http://www.fetalmedicineusa.com/accreditation.php (2 November 2011, date last accessed).
15. Pandya PP, Altman DG, Brizot ML *et al*. Repeatability of measurement of fetal nuchal translucency thickness. *Ultrasound Obstet Gynecol* 1995;**5**:334–337.
16. Bland M. *An Introduction to Medical Statistics*. 3rd edn. Oxford: Oxford University Press, 2000:248–9.
17. Wald A. Sequential tests of statistical hypotheses. *Ann Math Statist* 1945;**6**:117–56.

Appendix

R-SPRT

The R-SPRT, introduced by Grigg *et al*. [8], is a sequence of Wald SPRTs [17]. The cumulative score at the jth case is given by the iterative relationship $S_j = S_{j-1} + (X_j - s)$, where $0 < s < 1$, $S_0 = 0$, $X_j = 0$ if the jth case is a success (i.e. S decreases by s) and $X_j = 1$ if the jth case is a failure (i.e. S increases by $1-s$). Competency is achieved when the cumulative score crosses an absorbing (terminating) barrier at height h_0 ($h_0 < 0$). When the cumulative score crosses a resetting barrier at height h_1 ($h_1 > 0$), i.e. $S_j > h_1$, a trainee is considered incompetent and the cumulative score is reset to zero ($S_j = 0$).

The three variables s, h_0 and h_1 of an R-SPRT are defined by four test parameters: p_0 = acceptable failure rate, p_1 = unacceptable failure rate, α^* = Type I error rate parameter and β^* = Type II error rate parameter. The case score, s, is given by $s = Q/(P+Q)$, where $P = \ln(p_1/p_0)$ and $Q = \ln[(1-p_0)/(1-p_1)]$. The height of the terminating barrier is defined by $h_0 = -b/(P+Q)$, where $b = \ln[(1-\alpha^*)/\beta^*]$ and the height of the resetting barrier is defined by $h_1 = a/(P+Q)$, where $a = \ln[(1-\beta^*)/\alpha^*]$.

For a given trainee, over many cases, the mean increment per case of the cumulative score S is given by a value δ. If the failure probability for a single test is r, then the expected value of δ is given by the probability of success × increment for success + probability of failure × increment for failure (i.e. $\delta = (1-r)(-s) + r(1-s)$).

There exists a value of r for which the mean increment per case of the cumulative score is zero. From the definition of δ, this occurs when r is equal to the case score, s (i.e. when $r = s$, $\delta = 0$). If the failure probability for a single test exceeds s (i.e. $r > s$), then $\delta > 0$, and the cumulative score, S, increases on average and tends towards the resetting barrier. If the failure probability of a single test is $<s$ (i.e. $r < s$), then $\delta < 0$ and the cumulative score, S, decreases on average and tends towards the terminating barrier. If r is equal to the threshold value s (i.e. $\delta = 0$), the cumulative score neither increases nor decreases on average and will not tend towards either the terminating or the resetting barrier.

LC-CUSUM

In the LC-CUSUM, introduced by Biau *et al*. [2], the cumulative score at the jth case is defined by the iterative relationship $S_j = \min(0, S_{j-1} - W_j)$, where $S_0 = 0$ and W_j is a weight that depends on whether the jth case was a success ($W_j = W_s$, $W_s > 0$; i.e. S decreases) or failure ($W_j = W_f$, $W_f < 0$; i.e. S increases). Competency is achieved when the cumulative score crosses an absorbing (terminating) barrier at height h_0 ($h_0 < 0$). The definition of S_j implies the presence of a resetting barrier at height $h_1 = 0$.

The two weight values of an LC-CUSUM are defined by the acceptable failure rate (p_0) and the unacceptable failure rate (p_1), with weight for success, $W_S = \ln[(1-p_0)/(1-p_1)]$ and weight for failure, $W_f = \ln(p_0/p_1)$. Using the same notation as the R-SPRT, the weights can be expressed as $W_s = Q$ and $W_f = -P$.

The value of h_0 is the third parameter, which defines an LC-CUSUM test. It is a free parameter that, in practice, may be chosen through simulation to achieve the required outcome of a test. For this study, to permit direct comparison between R-SPRT and LC-CUSUM tests using the same p_0 and p_1 parameters, the value of h_0 for the LC-CUSUM tests was chosen to give the same minimum run length (i.e. number of successive successes needed to achieve competence, starting from the first case) as an R-SPRT with parameters (p_0, p_1, α^* and β^*). For an LC-CUSUM, this placed the height of the terminating barrier at $h_0 = \ln[(1-\alpha^*)/\beta^*]$.

For the LC-CUSUM, the mean increment per case of the cumulative score is given by $\delta = (1-r)(-Q) + rP$. As with the R-SPRT, there exists a value of the failure probability for a single test, r, for which the mean increment of S per case is zero. This occurs at $r = Q/(P+Q) = s$ at which the cumulative score tends towards neither barrier. If the failure probability for a single test is $<s$ (i.e. $r < s$), the cumulative score, S, decreases on average and tends towards the terminating barrier.