Clinical Radiology 74 (2019) 327.e7–327.e19

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Qualitative ultrasound training: defining the learning curve

P.J. Mullaney*

Department of Radiology, University Hospital of Wales, Cardiff, UK

ARTICLE INFORMATION

Article history:
Received 16 May 2018
Accepted 24 December 2018

AIM: To assess the developing competence of two trainees learning groin and shoulder ultrasound when compared to an expert practitioner.

MATERIALS AND METHODS: Specific pro formas were used to record ultrasound trainee performance in each scan region and their diagnosis was compared to the opinion of the expert. The data derived were reviewed using kappa analysis and training end points were defined as a minimum 80% trainee agreement with the expert. Retrospectively, cumulative sum analysis was applied to the data to assess case-by-case performance.

RESULTS: For groin hernias, reporting an average of 70 examinations was required to become competent and inguinal hernias required higher numbers of examinations than femoral hernias. For shoulders, an average of 80 examinations was required and the supraspinatus and infraspinatus tendons proved the most challenging structures.

CONCLUSIONS: Kappa analysis demonstrated a differential in the learning curves for individual structures within each examination region. Sequential kappa scores are consistent with a sigmoid learning curve. The numbers required to achieve satisfactory agreement are suggested as required minima for ultrasound training curricula. Cumulative sum analysis provided a sensitive indicator of trainee performance, quickly highlighting individual learning difficulties when they arose. Its prospective use can ensure extra training support is instigated quickly and appropriately.

Introduction

Ultrasound (US) training is a significant component of routine practice, which mandates considerable time input from senior staff. It is important to ensure trainees progress appropriately during training to justify the resources used, and to determine final competence.

Most US training curricula emphasise the numbers of examinations that must be performed to be considered competent,[1–4] but are less specific in assessing the *quality* of experience gained. There are various methods of assessment that can be adapted to provide qualitative and more objective data on performance. Direct comparison against a recognised reference standard provides the most robust data allowing sensitivity and specificity to be calculated, but these data are rarely available during practice and must often be applied retrospectively. Other methods include, but are not limited to, correlation coefficients, kappa statistics for inter-rater agreement, and cumulative sum analysis for case-by-case performance. The importance of objective tools in training assessment has been discussed[5,6]; however, there is little literature on the practical

* Guarantor and correspondent: PJ Mullaney, Department of Radiology, University Hospital of Wales, Cardiff, UK. Tel.: +44 (0)29 20743956.
E-mail address: peter.mullaney@wales.nhs.uk

https://doi.org/10.1016/j.crad.2018.12.018

implementation of such methods and the impact of their results. This prospective study was performed during the US training of sonographers learning new US skills in groin and shoulder US. The data derived were used to assess sonographer competence in qualitative terms and to provide an evidence base to justify independent practice. The data are presented using different methods of assessment and are discussed. The methods used in this study are suggested as models for defining US competence on more specific terms than are currently available.

Materials and methods

Two qualified sonographers with 11 and 17 years of experience in abdominopelvic and obstetric US were recruited to learn US examination of shoulders and groin hernias in response to increasing service demands. Neither sonographer had any previous experience of shoulder or groin hernia ultrasound. Both sonographers attended regular US clinics with an appropriate case mix (four groin examinations, six shoulder examinations). Training was performed by a single consultant musculoskeletal radiologist (P.J.M.). Ethics Board review was not required as the training forms part of normal practice in the institution. Patients were informed of this and their permission was obtained in all cases. Initial training involved witnessing examinations performed by the consultant and attending lectures on shoulder and groin anatomy/pathology, which are provided as part of a postgraduate diploma course on Musculoskeletal Ultrasound offered by the local affiliated university. This course has approval from the institution's Postgraduate Deanery and the Consortium for Accreditation of Sonographic Education (CASE). The sonographers then examined patients under the direct supervision of the consultant with feedback. Each sonographer witnessed or examined a total of 250 patients fulfilling training recommendations from the UK Royal College of Radiologists.[1]

With recommended training numbers accrued, trainee competence was then assessed prospectively, in a blinded manner using a minimum of 50 consecutive shoulders and groins for each trainee. The trainee examined the patient and their findings were recorded on a specific pro forma sheet prior to the consultant's examination (Electronic Supplementary Material, Appendix A). The trainee's diagnoses were then compared to the consultant's final opinion, and feedback was provided to the trainee after each case. For groin examinations, Cohen's kappa statistic[7] was used to assess inter-rater agreement on the presence or absence of a hernia. For shoulders, each of six structures—long head of biceps tendon (Bcps), subscapularis tendon (Subscap), supraspinatus tendon (SS), infraspinatus tendon (IS), subacromial bursa (Bursa), and acromioclavicular joint (ACJ)—required specific assessment. Weighted kappa was used to calculate agreement between the trainee and expert for each of the 4 main shoulder tendons and the ACJ using the technique described by Zaiontz.[8] Categories used for weighted kappa analysis for each tendon were: tendon normal, abnormal tendon not torn, or tendon torn. Where multiple pathologies were present, tendon tear was considered the most significant finding—as a potential indicator for surgery—and used for the analysis. Tendinosis, tenosynovitis, and calcification were all considered under the category of abnormal tendon not torn as these conditions tend to be managed actively but non-operatively. For the ACJ, categories defined were: structure normal or mild osteoarthritis (OA) present, moderate OA present, or severe OA present. For the Bursa, non-weighted kappa was calculated for the binary assessment of bursa normal or bursa abnormal. For all kappa calculations, rolling cohorts of 30 consecutive patients were used (i.e. Cohort 1: patients 1–30 inclusive, cohort 2: patients 2–31 inclusive, cohort 3: patients 3–32, etc.) and a minimum requisite score of 0.8 was set to define competence. The assessment period was initially set to run until this score was achieved by both trainees in each region examined.

Retrospectively, cumulative sum (CUSUM) performance analysis was calculated for each trainee in each region. CUSUM curves were derived using the methods described by Noyez.[9] Operation success was defined as diagnostic agreement with the expert and set a score of +0.2, operation failure was scored at −0.8. Satisfactory and unsatisfactory performance limits were calculated as described by Williams *et al.*[10] (Electronic Supplementary Material, Appendix B). Patient cohorts for each trainee in each region were compared for age (using the Student's *t*-test), the side of the body examined, and the referral source (both using Fisher's exact test). Calculations were performed using Microsoft Excel (Microsoft Corporation, Seattle, WA, USA). Statistical analysis was performed using Graphpad Instat (GraphPad Software, Inc. La Jolla, CA, USA).

Results

The training period ran from April 2009 to February 2013. The trainees were trained simultaneously, but competing service demands would cause the trainees to miss sessions sporadically. It therefore took the trainees different lengths of time to accrue patient numbers sufficient to demonstrate satisfactory performance.

Hernias

Trainee 1 was assessed over 81 patients, trainee 2 over 71 patients. Patient numbers and characteristics for each

Table 1
Patient cohort characteristics for each trainee.

Trainee 1	Trainee 2
81 Groins	71 Groins
Average age: 54.2 years (SD 17.7 yrs)	Average age: 52 years (SD 16.4yrs)
44 right, 37 left	35 right, 36 left
Referral source: 1 gynaecology, 1 dermatology, 1 renal transplant, 14 surgery, 64 general practice	Referral source: 1 urology, 1 rheumatology, 7 surgery, 62 general practice

No significant difference in age (Student's t p=0.4311), right/left distribution (Fisher's exact test p=0.6258), or referral source (Fisher's exact test p=0.2406) between the cohorts for each trainee.

trainee are presented in Table 1. The number of patients examined to achieve required kappa scores are presented in Table 2. Kappa curves for inguinal hernia (a) and femoral hernia (b) diagnosis for each trainee are presented in Fig 1. For trainee 1, the kappa curve for inguinal hernia demonstrates a generally consistent upward trajectory achieving a final kappa score of 0.93 at patient 79 (cohort 50). For trainee 2, the inguinal hernia kappa curve is initially downward sloping—kappa scores decreasing from 0.73 at patient 30 (cohort 1) to 0.66 at patient 56 (cohort 27)—but thereafter a consistent upward gradient is demonstrated with a final kappa score of 0.93 at patient 71 (cohort 42).

For femoral hernias, trainee 1 demonstrates a flat trajectory until patient 41 (cohort 12, k=0.63). Kappa increases to 0.9 by patient 52 (cohort 23). Thereafter the kappa curve is relatively flat, with kappa ≥0.8 up to patient 76 (cohort 47). For the final five patients, kappa decreases slightly (final kappa score: 0.73). Trainee 2 also demonstrates an initial flat trajectory up to patient 48 (cohort 19), but thereafter a consistent positive gradient for the remainder of the training period with a final kappa score of 0.84 by patient 71 (cohort 42).

CUSUM plots for each trainee for inguinal (a) and femoral hernias (b) are presented in Fig 2. For inguinal hernias, both trainees demonstrate relatively flat performance curves. Although the curve for trainee 1 falls below zero between patients 2 and 40, performance remains within acceptable performance limits. From patient 33 onwards, performance improves, with more consistent operation success, crossing a positive performance threshold at patient 75. Trainee 2 achieves similar consistency from patient 28, crossing a positive performance threshold at patient 56. For femoral hernias, both trainees achieve consistent operation success almost immediately, crossing two positive performance thresholds during training.

Shoulders

Trainee 1 was assessed over 117 patients, trainee 2 over 86 patients. Patient cohort characteristics for each trainee are summarised in Table 3. The number of patients required to achieve minimum required kappa scores are presented in Table 4. Results are presented by each structure examined.

Bcps (Fig 3)

Trainee 1 immediately and consistently demonstrates performance above required minimum standards. For all patients after patient 35 (Cohort 6) kappa exceeds 0.8 and does not fall below the minimum required standard. For trainee 2, kappa lies close to zero until patient 69 (cohort 40) and does not achieve minimum required standards until patient 72 (Cohort 43, k=0.88). Once satisfactory performance has been achieved by trainee 2, it is maintained for the rest of the training period. CUSUM trends are more closely aligned between trainees demonstrating consistent operation success from the start and crossing multiple positive performance thresholds.

Table 2
Patient numbers (cohort number) required to achieve a kappa value of 0.8 or more for each trainee with mean figures where possible.

Hernia type	Trainee 1	Trainee 2	Mean
Inguinal	76 (47)	65 (36)	71
Femoral	47 (18)	69 (40)	58

Subscap (Fig 4)

For Trainee 1, kappa scores initially decrease from 0.75 at patient 30 (cohort 1) to 0.35 at patient 81 (Cohort 52). Thereafter a rapid increase in kappa is observed and maintained for the rest of the training period. Trainee 2 achieves satisfactory performance by patient 51 (Cohort 22), but performance deteriorates slightly by patient 57 (Cohort 28) and scores between 0.62 and 0.7 are maintained for the rest of the training period.

The CUSUM curve for trainee 1 is relatively flat but within acceptable limits up to patient 58. Thereafter consistent operation success is achieved and two positive performance thresholds are crossed by the end of training. For trainee 2, performance is below expected until patient 20 but still acceptable. From patient 14 to patient 67, consistent success is achieved and a positive performance threshold is crossed. For the final 15 patients, trainee 2's performance deteriorates slightly but remains acceptable.

SS (Fig 5)

Both trainee's kappa curves are initially flat with more consistent upward gradients in the latter half of training. For CUSUM values, trainee 1 demonstrates satisfactory performance up to patient 81 and improves thereafter crossing one positive performance threshold. Trainee 2 demonstrates poor performance for the first 34 patients with consistent failures and performance crosses a negative performance threshold at patient 22, and again at patient 32. From patient 34 onwards performance improves rapidly: performance limits are re-crossed upwards at patient 50 and an upward trajectory is maintained thereafter.

IS (Fig 6)

For trainee 1, agreement is no better than chance until patient 64 (Cohort 35, k=0.14), thereafter a rapid increase in kappa is seen until a maximum kappa of 0.94 is achieved by patient 100 (Cohort 71). Kappa then deteriorates for the final 18 patients, with a final kappa score of 0.56. Trainee 2's kappa curve demonstrates a consistent increase up to patient 77 (cohort 48). Although there is a slight decrease in performance thereafter, kappa scores remain satisfactory (final k=0.79). CUSUM scores for Trainee 1 demonstrates inconsistent but acceptable performance up to patient 47. For the next 40 patients, performance improves with consistent success crossing one positive performance threshold. Thereafter regular failures occur and the performance threshold is re-crossed in a downwards direction by the end of training. CUSUM scores for trainee 2 demonstrate

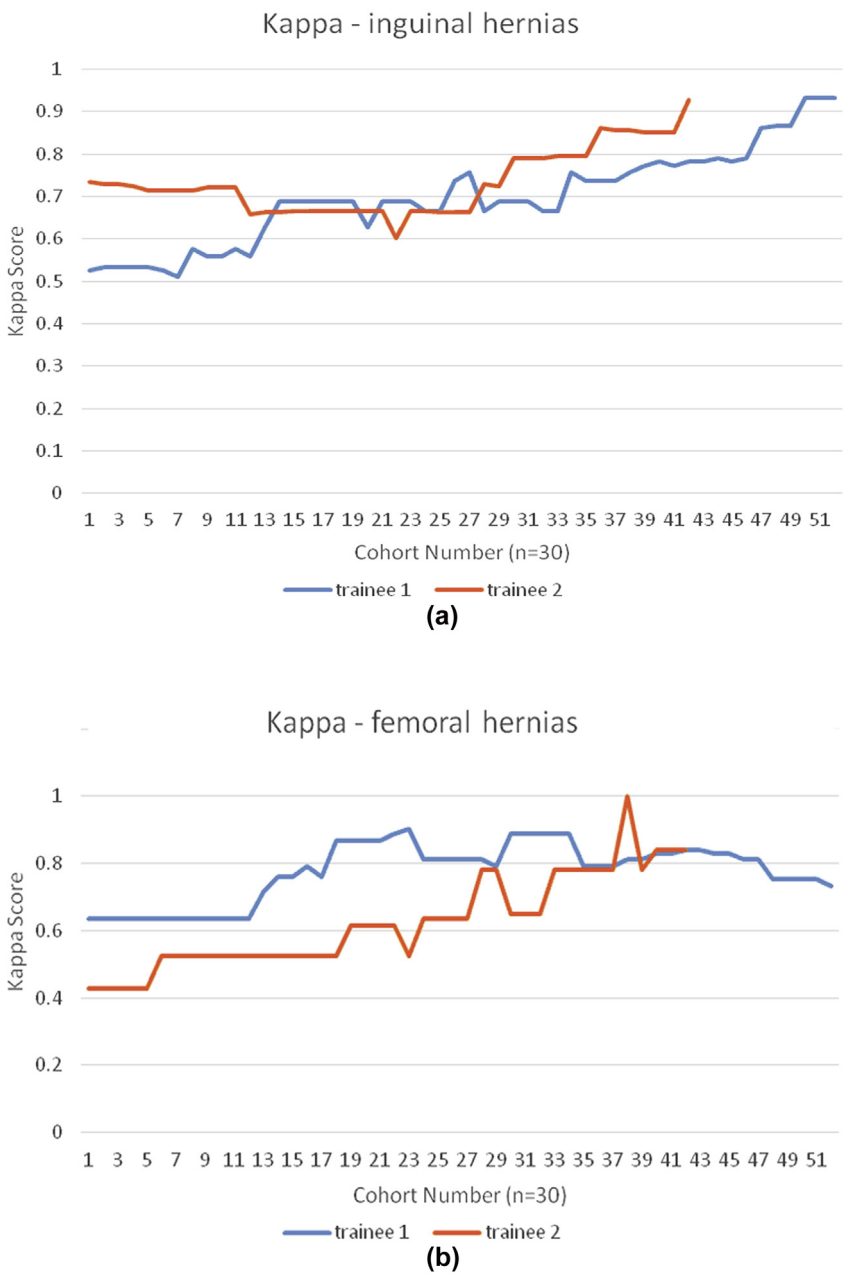


Figure 1 Kappa scores for both trainees for (a) inguinal hernias and (b) femoral hernias. Patient cohorts are defined in the Materials and Methods section.

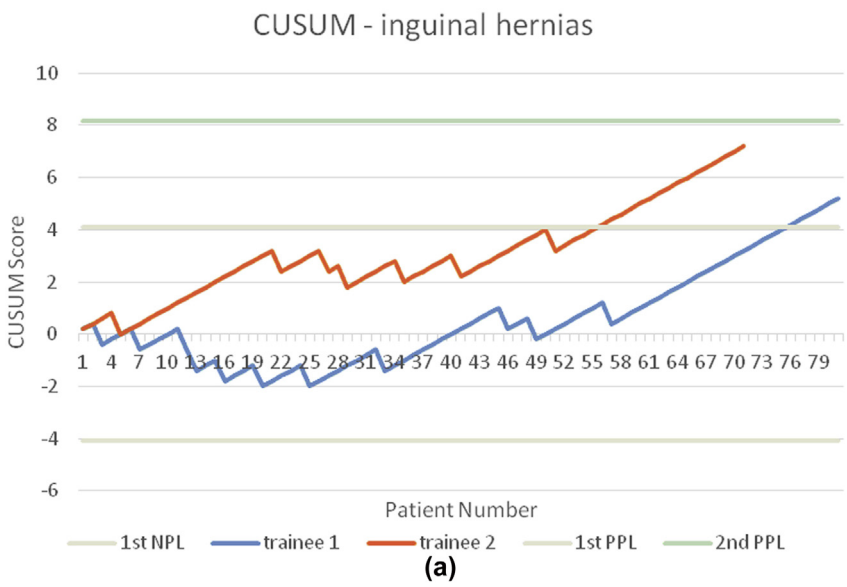
consistent operation success and three positive performance thresholds are crossed by the end of training.

ACJ (Fig 7)

Both trainees demonstrate an initial flat kappa curve with a period of improvement later in training. Thereafter both trainees' performance deteriorates (see discussion). The CUSUM plots for both trainees demonstrate inconsistent but generally satisfactory performance for the first half of training. Trainee 1 demonstrates a period of consistent success from patient 60 to patient 96 crossing a positive performance threshold but performance deteriorates thereafter re-crossing this threshold downwards at patient 103. Trainee 2's performance is inconsistent but generally satisfactory throughout.

Bursa (Fig 8)

Both kappa curves demonstrate an upward trajectory after an initial flat/downward phase which for trainee 1 lasts until patient 60 (cohort 31), for trainee 2 patient 74 (cohort 45). Trainee 1's performance deteriorates for the final 26 patients (k=0.93 at cohort 62 to 0.73 at cohort 88). The CUSUM plot for trainee 1 is satisfactory up to patient 57, with consistent success thereafter, crossing one performance threshold upwards. For trainee 2, performance is unsatisfactory and two negative performance thresholds are crossed by patient 47.



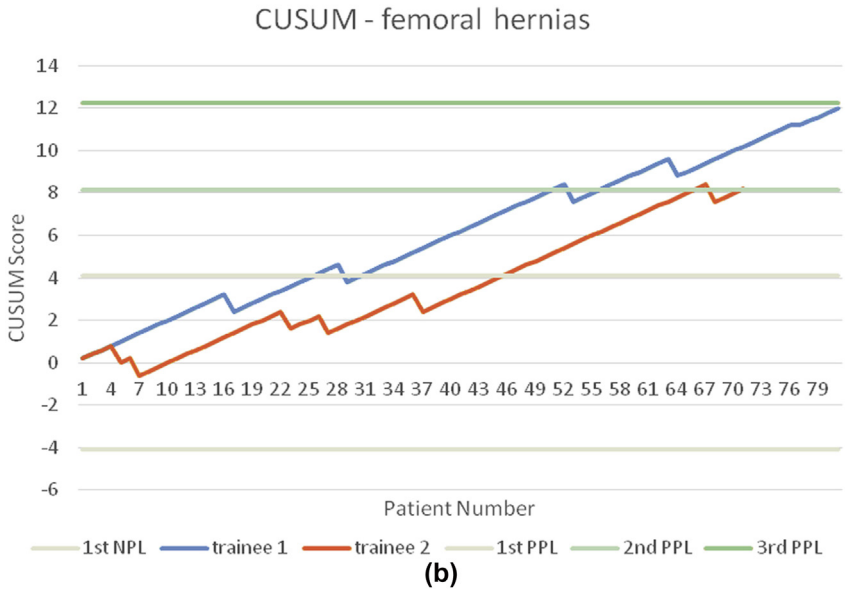


Figure 2 CUSUM scores for both trainees for (a) inguinal hernias and (b) femoral hernias. NPL, negative performance limit; PPL, positive performance limit. See Electronic Supplementary Material, Appendix B.

Table 3
Patient cohort characteristics for each trainee.

Trainee 1	Trainee 2
117 Shoulders	86 shoulders
Average age 54.9 year (SD 14.9 years)	Average age 55 years (SD 15.2 years)
58 right, 59 left	54 right, 32 left
2 shoulders post rotator cuff repair	1 shoulder post hemiarthroplasty
Referral source: 3 physiotherapy, 8 rheumatology, 12 surgery, 94 general practice	Referral source: 2 physiotherapy, 7 rheumatology, 11 surgery, 66 general practice

No significant difference in age (Student's t p=0.9621), right/left distribution (Fisher's exact test p=0.0653), or referral group (Fisher's exact test p=0.6559) between population groups for each trainee.

Table 4
Patient numbers required to achieve a kappa or weighed kappa value of 0.8 or more for each trainee with mean figures where possible.

Structure	Trainee 1	Trainee 2	Mean
Bcps	35	72	54
Subscap	86	51	69
SS	94	72	83
IS	82	72	77
ACJ	75	-	-
Bursa	84	-	-

Bcps, long head of biceps tendon; Subscap, subscapularis tendon; SS, supraspinatus tendon; IS, infraspinatus tendon; Bursa, subacromial bursa; ACJ, acromioclavicular joint.

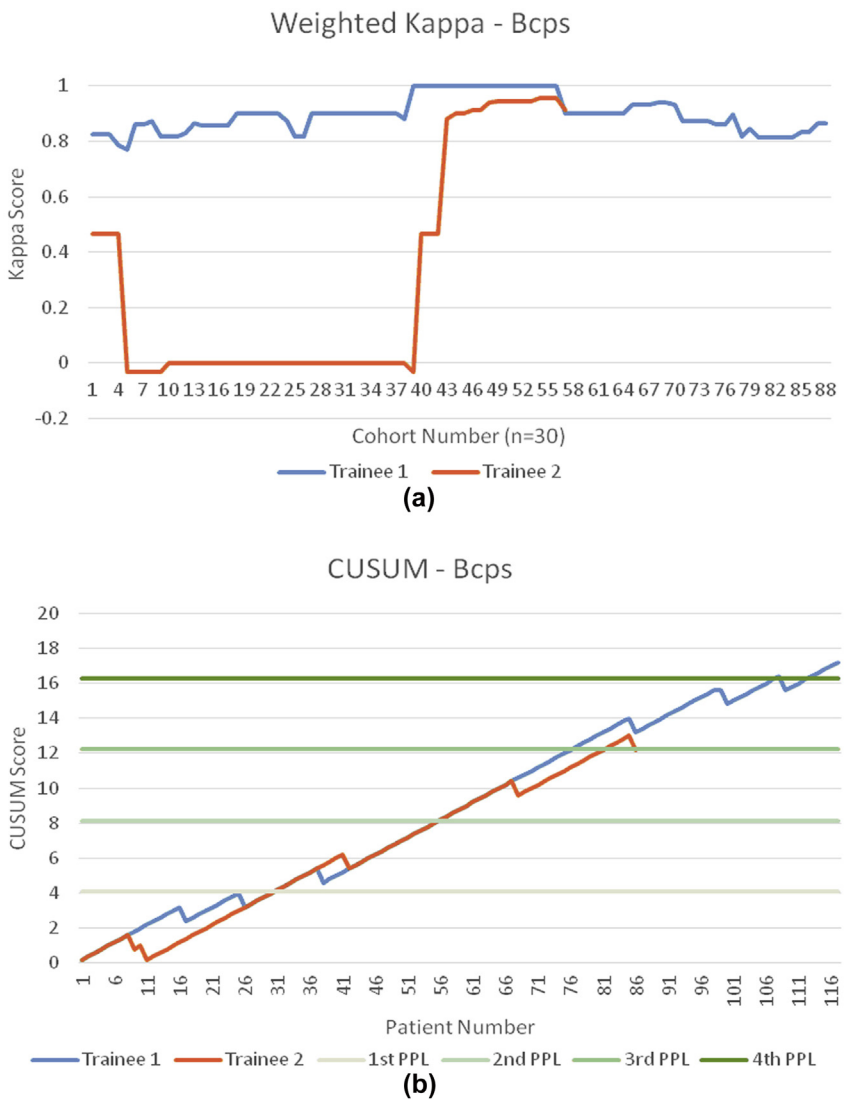


Figure 3 (a) Weighted kappa and (b) CUSUM plot for each trainee for the Bcps tendon. In (a) cohorts are defined in the Materials and Methods section. In (b) NPL, negative performance limit; PPL, positive performance limit.

Thereafter the downward trajectory ceases with improved performance for the final 14 patients.

Discussion

Given the operator-dependant nature of US, effective training is vital to engender consistent, effective performance. Previous studies have attempted to quantify the numbers of US cases required to achieve a predetermined level of competence in emergency examination,[11] US-guided endotracheal intubation,[12] transvaginal US training using a simulator[13] and US-guided jugular venous cannulation,[14] but none describe practical methods of prospective assessment.

Of recognised learning models, both exponential learning curves and sigmoid learning curves are described (Fig 9). They are encountered empirically[15,16] and can be modelled mathematically.[17] With exponential curves a failure-based learning model is used to describe a steep (usually upward) trajectory early in the training period, reflecting rapid increases in skill with increasing experience. For sigmoid curves, a success-based model is used: there is a flat or shallow upward trajectory early in the training period and the steeper phase of rapid skill acquisition is encountered later. The acquisition of task-irrelevant skills (i.e., familiarisation with necessary equipment) may explain this. The phase of rapid skill increase is sometimes termed the "eureka" moment where new insights and positive feedback facilitate rapid progress. In the later phases of training, both curves demonstrate asymptotic flattening as further increases in skill require a disproportionate amount of experience to achieve.

As part of this study, two assessment pro formas are presented which allow a trainee's diagnostic competence for shoulder and groin US (Electronic Supplementary Material, Appendix A) to be recorded and compared against an expert in more detail than is currently used. For hernias, a schematic allows the operator to record information on the

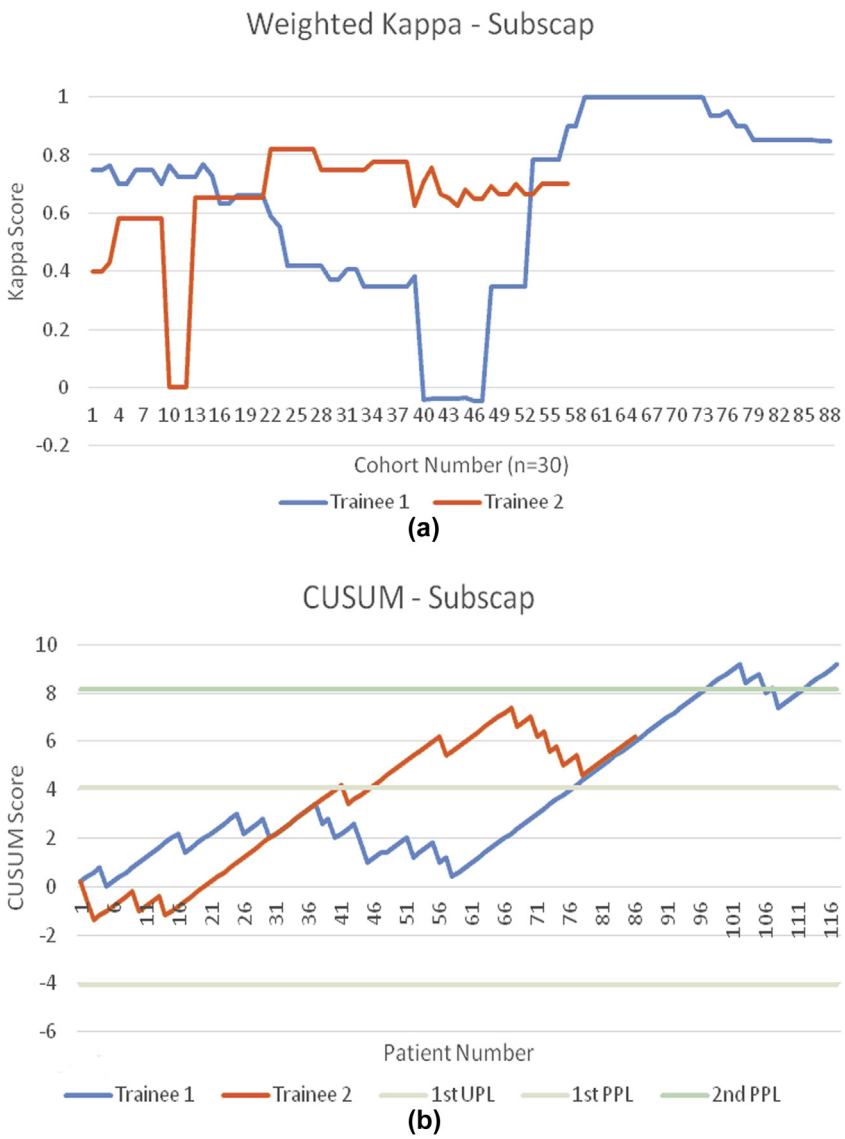


Figure 4 (a) Weighted kappa and (b) CUSUM plot for each trainee for the Subscap tendon. In (a) cohorts are defined in the Materials and Methods section. In (b) NPL, negative performance limit; PPL, positive performance limit.

type, size, and contents of inguinal hernias which can be applied to various classification systems. Any femoral hernias must also be recorded. Previous surgery and the presence of any indwelling mesh can be noted. The shoulder pro forma requires a separate diagnosis for each of six clinically relevant structures. A schematic of the SS tendon in the axial plane is divided into six regions allowing the trainee to localise tendon pathology and define its extent. A comments section allows confounding factors (e.g., post-operative tendon, limited range of movement, high body mass index) to be acknowledged. The ACJ is assessed for severity of osteoarthritis and active inflammation. Features indicating subacromial bursitis can be documented for future review. These pro formas have been used successfully by various trainees including sonographers, radiology trainees and an academic physiotherapist. The shoulder pro forma has been the model for a "training package" for shoulder US.[18] The information provided by these pro formas can highlight specific performance issues where extra training and support may be required. Conversely, they may support an earlier completion of training once satisfactory diagnostic performance has been demonstrated over an agreed number of consecutive cases. The principles used to create these pro formas could be applied to other (musculoskeletal) regions to create a comprehensive training portfolio across a whole syllabus.

Cohen's kappa statistic and CUSUM analysis have been used to analyse the pro forma training data: Cohen's kappa statistic was formulated to quantify the level of inter-rater agreement, correcting for chance agreement between raters.[7] It is frequently used in medical studies, but its results should be interpreted with caution: kappa values can be distorted in small data sets or where the prevalence of the required diagnostic criterion is either very low or very high.[19] Assumptions made calculating the kappa statistic from the contingency tables can also be a cause of artefacts

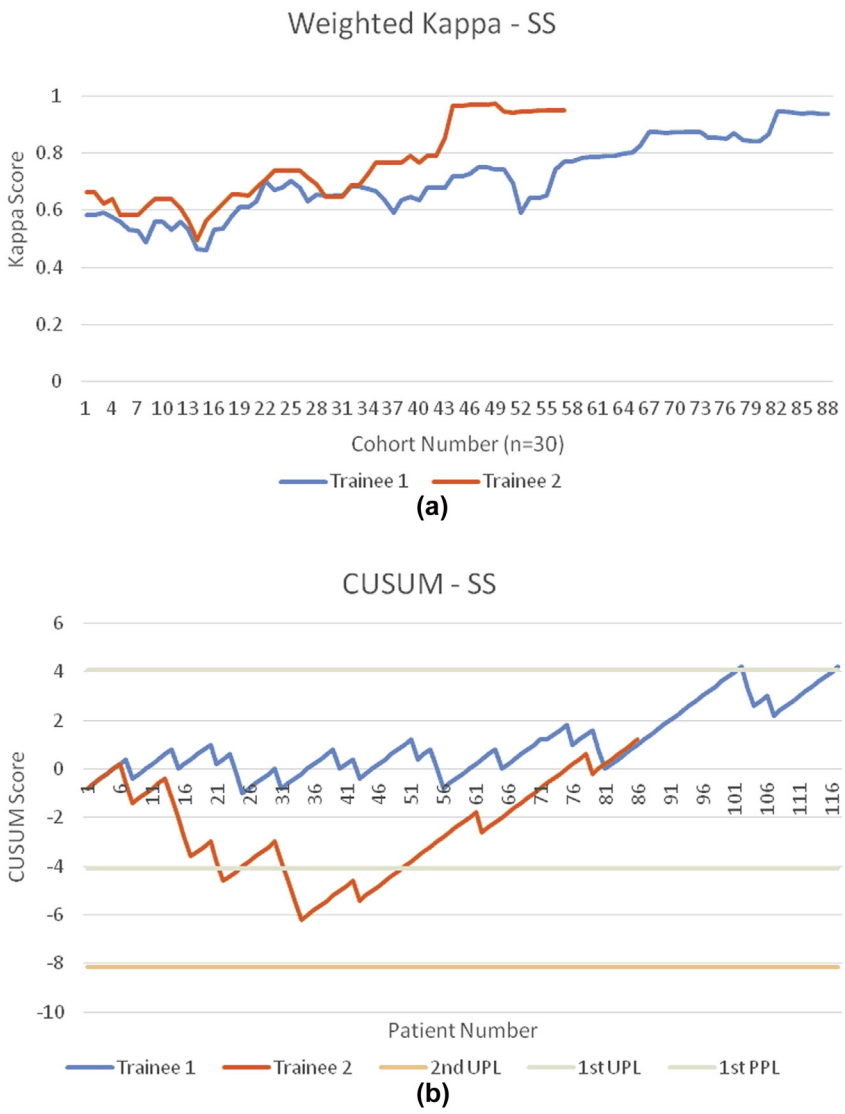


Figure 5 (a) Weighted kappa and (b) CUSUM plot for each trainee for the SS. In (a) cohorts are defined in the Materials and Methods section. In (b) NPL, negative performance limit; PPL, positive performance limit.

leading to paradoxical results.[20–22] Finally, the accepted scale for grading agreement is arbitrary.[23] Where a kappa value of 0.6 or more has been classed as "good agreement", this still allows discrepancy between observers of up to 40%. In this study, cohorts of 30 patients are used in to reduce artefacts from small data sets; other sources of bias are discussed where relevant. Weighted kappa quantifies inter-rater agreement where there are more than two possible diagnoses. Here the technique described by Zaiontz[8] is used for the shoulder tendons and the ACJ. Nonlinear weighting has been used to emphasise the clinical importance of rotator cuff tears (Electronic Supplementary Material, Appendix B), but weighting can be adjusted by any criteria considered relevant. For the Bursa, using non-weighted kappa seems practical as most abnormal bursae are treated with a targeted injection.

CUSUM curves were initially formulated to monitor manufacturing quality in the munitions industry, but they have since been used effectively to monitor performance in such disciplines as cardiac surgery,[24,25] vascular intervention,[26] anaesthetics,[27] and endoscopy.[28] Noyez[9] describes this technique in detail but briefly: a procedure is selected for review and operation success or failure are both defined and assigned a value (e.g. positive for success and negative for failure). The cumulative sum of these values across a number of consecutive operations can then be derived and plotted on a chart. The gradient and overall height of the plot so derived provides feedback on performance on a case by case basis: operation failure will produce a negative slope and lower the overall score. If this cumulative score crosses a predetermined limit, an intervention to address performance is required. The utility of CUSUM is in determining when operation failures are within acceptable limits of variation (in control) or fall outside (out of control), and when to intervene appropriately. Performance limits can also be dynamically adjusted over training[29] but the

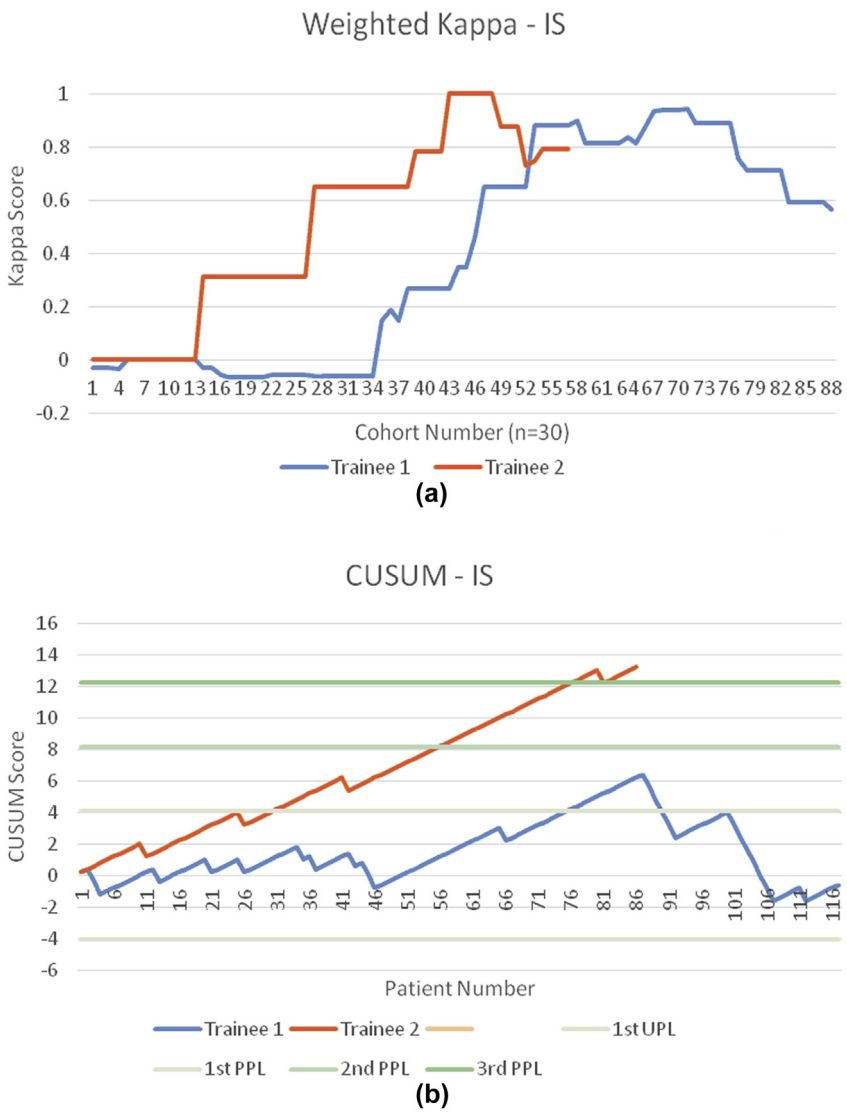


Figure 6 (a) Weighted kappa and (b) CUSUM plot for each trainee for the IS. In (a) cohorts are defined in the Materials and Methods section. In (b) NPL, negative performance limit; PPL, positive performance limit.

calculations required to set them are complex and can require specialist statistics input. In this study, satisfactory and unsatisfactory performance limits are calculated as described by Williams *et al.*[10] (Electronic Supplementary Material, Appendix B). Going forward, relevant action limits could be set using centile scores from pooled trainee performance data. Although performance limits above the zero line are "good" and those below "bad", the direction that any performance limit is crossed also has relevance: crossing any performance limit in the upward direction denotes consistent operation success and vice versa. Although a curve below the zero mark denotes a higher proportion of operation failures than expected, the curve's trajectory should also be considered. An example is the CUSUM plot for trainee 2 examining the SS (Fig 5b): after patient 34, trainee 2 achieves consistent operation success and performance thresholds are crossed in an upward direction. Similarly, crossing performance threshold(s) in a downward direction merits attention, regardless of overall score. This is demonstrated by trainee 1 examining the IS (Fig 6b): from patient 87 onwards performance limits are crossed in a downward direction although overall scores remain above zero.

The results presented demonstrate the learning curve for two trainee operators new to the technical demands of musculoskeletal US. Although the trainees were experienced in abdominopelvic and obstetric US, they found this experience was not transferable. All aspects of musculoskeletal US, from the handling of the transducer, relevant anatomy, anisotropy, patient positioning, and coordinating Valsalva manoeuvres presented novel challenges. A sigmoid learning profile might therefore be expected in their kappa curves as they contend with these new skills. For hernias, the kappa scores of both trainees tend to rise more quickly in the latter half of training and demonstrate a relatively static early phase, which is more consistent with a sigmoid learning curve than an exponential one. Kappa scores suggest that approximately 20% more cases are required to

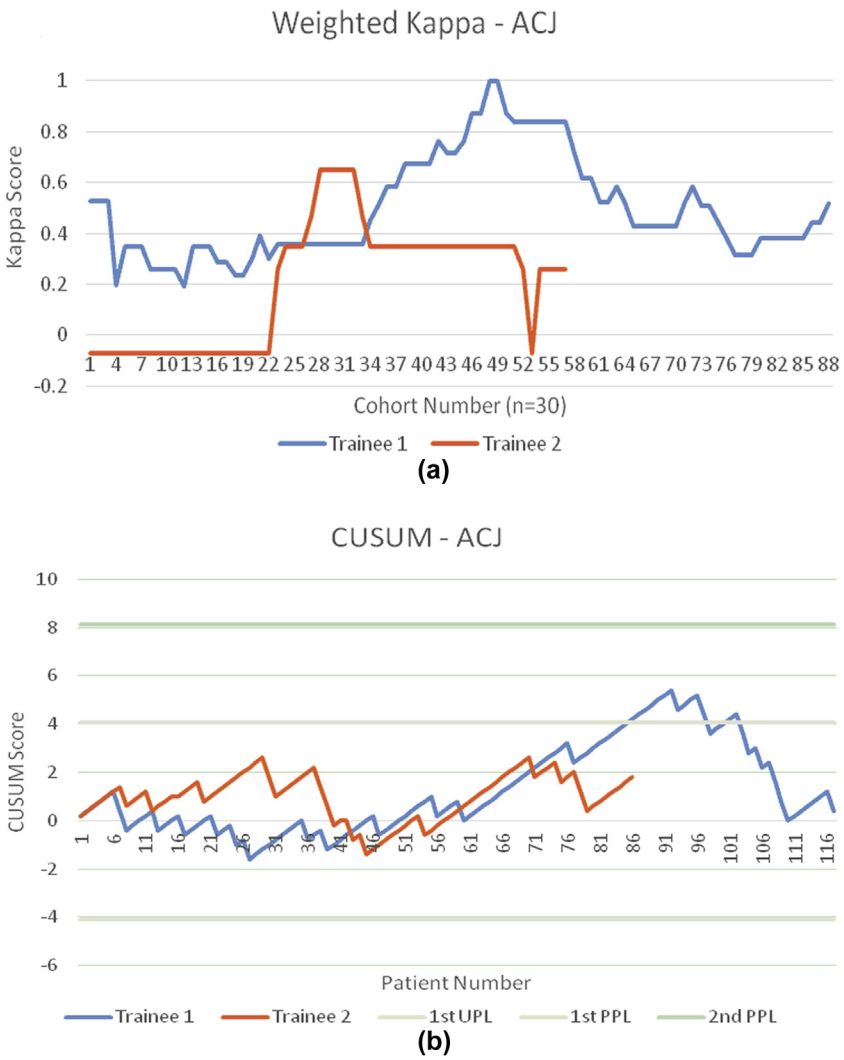


Figure 7 (a) Weighted kappa and (b) CUSUM plot for each trainee for the ACJ. In (a) cohorts are defined in the Materials and Methods section. In (b) NPL, negative performance limit; PPL, positive performance limit.

achieve competence diagnosing inguinal hernias over femoral hernias (Table 2). This could be due to the more complex anatomy of the inguinal canal over the femoral canal and greater susceptibility to confounding factors such as obesity and previous surgery.

For shoulder US, there is an appreciable similarity to the overall shape of the structure specific kappa curves between trainees, although individual kappa scores and patient numbers differ. An exception to this similarity is seen in the Bcps tendon (Fig 3a). Here trainee 1 achieves satisfactory kappa scores almost immediately and maintains this performance throughout. For trainee 2, kappa scores demonstrate agreement no greater than chance for the majority of training. Prevalence bias probably explains this discrepancy: for trainee 2, only two of the first 67 Bcps tendons were considered abnormal by the expert (2.9% "prevalence") versus six out of the following 19 patients thereafter (31.6% "prevalence"). For trainee 1, 11 out of the first 67 patients were judged to have Bcps pathology by the expert (16.4% "prevalence"). The Bcps CUSUM plots (Fig 3b) support this, with no significant discrepancy in performance between trainees over the whole of the training period. Where true learning difficulties arose, differentiating Bcps tendon dislocation from tear proved to be the most challenging issue.

The kappa curves for most shoulder structures are again more easily reconciled to a sigmoid learning curve consistent with a success-based learning paradigm. For several structures kappa scores tend to plateau toward the end of training. Exceptions to these trends are seen in the ACJ and IS where scores deteriorate significantly. For the ACJ, the criteria used to judge the presence and severity of ACJ osteoarthritis were adjusted in the latter half of training at the request of orthopaedic surgeons. Although these changes were communicated to trainees, it is likely that accommodating these changes impacted performance, both of the trainees *and* the expert. This deterioration in kappa scores occurred at a similar time for both trainees supporting this conclusion. For IS, there is no clear cause for deteriorating performance and it is assumed that, even

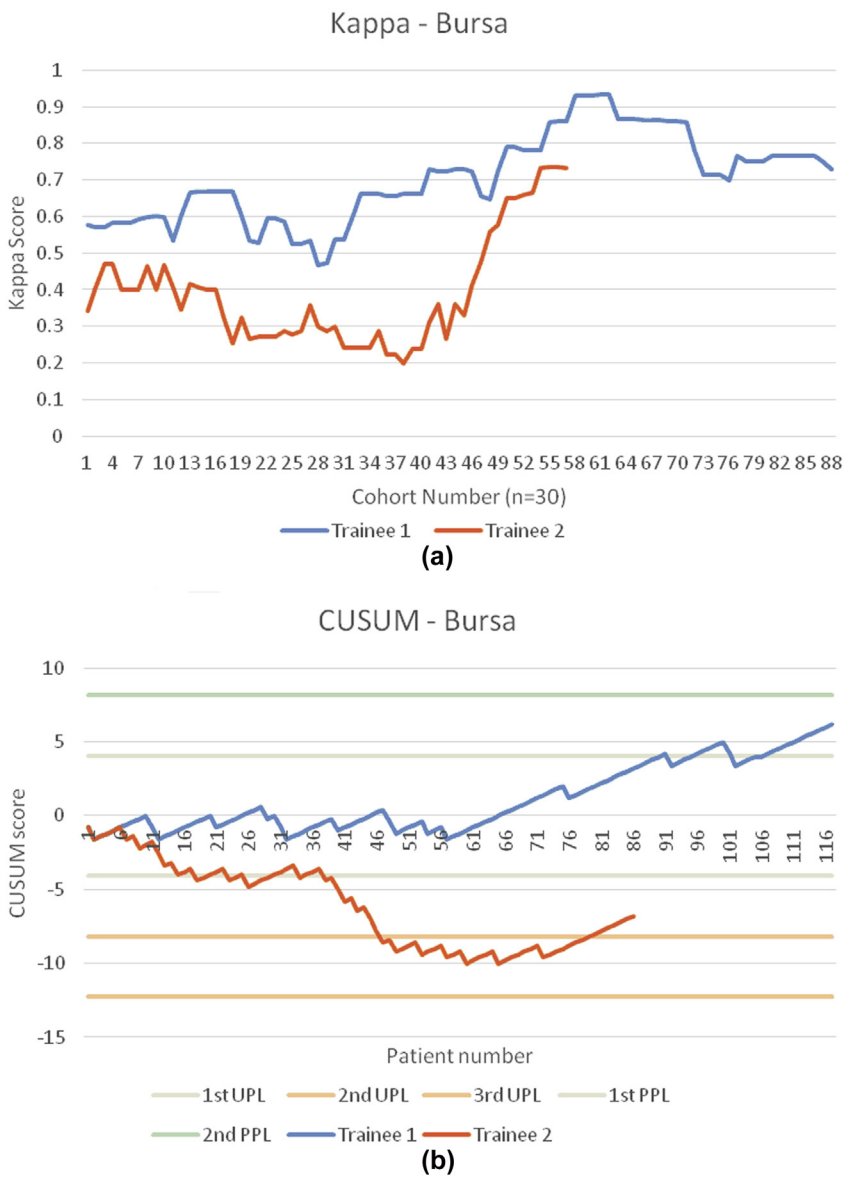


Figure 8 (a) Weighted kappa and (b) CUSUM plot for each trainee for the Bursa. In (a) cohorts are defined in the Materials and Methods section. In (b) NPL, negative performance limit; PPL, positive performance limit.

towards the end of training, trainee experience remains relatively limited allowing difficult cases to impact diagnostic confidence. Learning points for IS lay in recognising when IS was also involved in SS tears, differentiating true tendinopathy from anisotropy and in differentiating a completely torn IS tendon from teres minor.

Table 3 summarises the numbers of shoulder examinations required by each trainee to achieve 80% or higher agreement with the expert examiner. It can be seen that the SS and IS tendons require between 20% and 50% more experience than the Subscaps and Bcps tendons to achieve satisfactory performance. The CUSUM plots broadly support this when the numbers of patients required to achieve "consistent" operation success are considered. The clinical emphasis on SS pathology as a determinant for surgery, coupled with the technical demands of tendon anisotropy and patient positioning may explain this prolonged learning curve.[30] In addition, accurately determining the size and location of incomplete tears may contribute to the need for greater experience to attain competence, although this was not specifically assessed for this study. When comparing trainee performance, trainee 2 appears to learn faster than trainee 1, achieving required kappa scores for the major tendons earlier than trainee 1. These trends support the overall impression of the trainer that trainee 2 was a more confident student, requiring less reassurance in difficult cases. These data could be used to modify the training period for such trainees allowing resources to be allocated more appropriately.

In summary, this study has demonstrated the utility of examination pro formas when assessing the performance of US trainees. The information derived can be used to assess trainee progress with a variety of methods. Kappa scores monitor overall trainee progress and provide evidence of

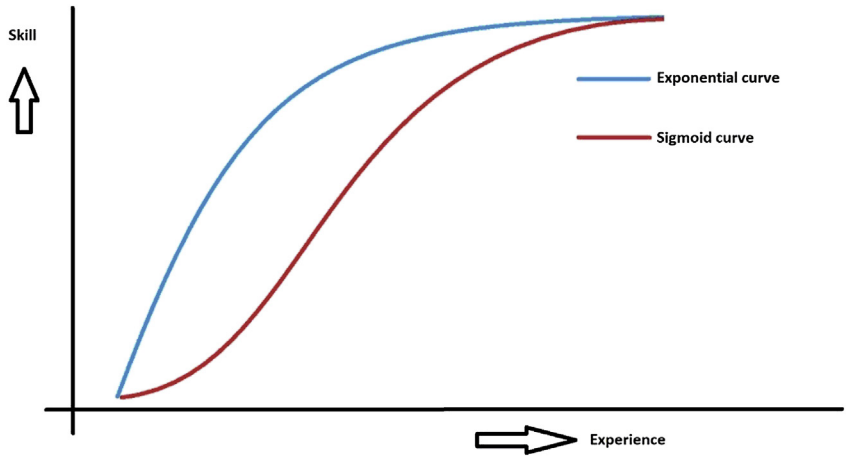


Figure 9 Schematic describing two common learning curve models. Both are characterised by a phase of rapid learning causing a steep upward slope, followed by an asymptotic flattening as increases in skill require increasing amounts of experience to achieve. For sigmoid curves, there is an initial flat or shallow upward phase before the phase of rapid learning.

competence at the end of training. CUSUM analysis allows a sensitive *prospective* assessment of trainee progress allowing extra support to be provided quickly and appropriately when learning issues become evident. Comparing information from both methods can mitigate statistical artefacts which might negatively impact trainee assessment. For groin hernias, a minimum of 70 examinations is suggested to attain competence, with inguinal hernias being the most difficult to master. For shoulder US, SS and IS appear to be the most difficult structures to assess reliably, suggesting a minimum of 80 examinations as a training requirement. Although the conclusions derived from only two students and two specific musculoskeletal regions must be limited, these figures are similar to a large retrospective review.[11] These assessment techniques are currently being applied to more US trainees, and generalised to trainee reporting in other imaging modalities to increase data available for analysis. It is emphasised that competence should only be awarded when *all* structures in a region can be reliably assessed. Going forward, pooled trainee performance data could be used to set more realistic logbook requirements and inform future US training curricula.

Conflict of interest

The author declares that there are no conflicts of interest for any part of this study.

Acknowledgements

The author thanks Dr Mark Kelson PhD, formerly of the South East Wales Trials Unit, Cardiff University for advice on kappa statistics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crad.2018.12.018.

References

1. The Royal College of Radiologists. *Ultrasound training recommendations for medical and surgical specialities*. Second Edition. London: The Royal College of Radiologists; 2012.
2. Education and Practice Standards Committee. European Federation of Societies for ultrasound in medicine and biology. Minimum training requirements for the practice of medical ultrasound in Europe. *Ultraschall Med* 2006;**27**:79–105.
3. The Royal College of Radiologists. *Focused ultrasound training standards*. London: The Royal College of Radiologists; 2012.
4. Level 2 Working Group, College of Emergency Medicine Ultrasound Subcommittee. *Guidance for level 2 ultrasound practice in emergency medicine*. London: The College of Emergency Medicine; 2008.
5. Jaffer A, Bednarz B, Challacombe B, *et al*. The assessment of surgical competency in the UK. *Int J Surg* 2009;**7**:12–5.
6. Nix CM, Margarido CB, Awad IT, *et al*. A scoping review of the evidence for teaching of ultrasound-guided regional anesthesia. *Reg Anesth Pain Med* 2013;**38**:471–80.
7. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Assess* 1960;**20**:37–46.
8. Zaiontz C. *Real statistics using excel*. 2017. Available at: www.real-statistics.com/reliability/weighted-cohens-kappa/. [Accessed 14 December 2017].
9. Noyez L. Cumulative sum analysis: a simple and practical tool for monitoring and auditing clinical performance. *Health Care Curr Rev* 2013;**2**:113–5.
10. Williams SM, Parry BR, Schlup MMT. Quality control: an application of the cusum. *BMJ* 1992;**304**:1359–61.
11. Blehar DJ, Barton B, Gaspari RJ. Learning curves in emergency ultrasound education. *Acad Emerg Med* 2015;**22**:574–82.
12. Chenkin J, McCartney CJ, Jelic T, *et al*. Defining the learning curve of point-of-care ultrasound for confirming endotracheal tube placement by emergency physicians. *Crit Ultrasound J* 2015;**7**:1–7.
13. Madsen ME, Konge L, Nørgaard LN, *et al*. Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. *Ultrasound Obstet Gynecol* 2014;**44**:693–9.
14. Nguyen BV, Prat V, Vincent JL, *et al*. Determination of the learning curve for ultrasound-guided jugular central venous catheter placement. *Intensive Care Med* 2014;**40**:66–73.
15. Gallistel CR, Fairhurst S, Balsam P. The learning curve: implications of quantitative analysis. *PNAS* 2004;**101**:13124–31.
16. Ritter FE, Schooler LJ. The learning curve. In: Smelser NJ, Baltes PB, editors. *International encyclopaedia of the social and behavioural sciences*. Oxford: Pergamon; 2001. p. 8602–5.

17. Leibowitz N, Baum B, Enden G, *et al*. The exponential learning equation as a function of successive trials results in sigmoid performance. *J Math Psychol* 2010;**54**:338–40.
18. Smith MJ, Rogers A, Amso N, *et al*. A training, assessment and feedback package for the trainee shoulder sonographer. *Ultrasound* 2014:1–14.
19. Grove WM, Andreasen NC, McDonald-Scott P, *et al*. Reliability studies of psychiatric diagnosis. *Arch Gen Psychiatr* 1981;**38**:408–13.
20. Feinstein AR, Cicchetti DV. High agreement but low kappa. 1. A problem of 2 paradoxes. *J Clin Epidemiol* 1990;**43**:543–9.
21. Cicchetti DV, Feinstein AR. High agreement but low kappa. 2. Resolving the paradoxes. *J Clin Epidemiol* 1990;**43**:551–8.
22. Cicchetti DV. When diagnostic agreement is high, but reliability is low: some paradoxes occurring in joint independent neuropsychology assessment. *J Clin Exp Neuropsychol* 1988;**10**:605–22.
23. Landis RL, Kock GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
24. Novick RJ, Stitt LW. The learning curve of an academic cardiac surgeon: use of the CUSUM method. *J Card Surg* 1999;**14**:312–20.
25. Murzi M, Cerillo AG, Bevilacqua S, *et al*. Traversing the learning curve in minimally invasive heart valve surgery: a cumulative analysis of an individual surgeon's experience with a right mini-thoracotomy approach doe aortic valve replacement. *Eur J CardioThorac Surg* 2012;**41**:1242–6.
26. Forbes TL, DeRose G, Kribs SW, *et al*. Cumulative sum failure analysis of the learning curve with endovascular abdominal aortic aneurysm repair. *J Vasc Surg* 2004;**39**:102–8.
27. Starkie T, Drake EJ. Assessment of procedural skills training and performance in anaesthesia using cumulative sum analysis (Cusum). *Can J Anaesth* 2013;**60**:1228–39.
28. Waller HM, Connor SJ. Cumulative sum (Cusum) analysis provides an objective measure of competency during training in endoscopic retrograde cholangio-pancreatography (ERCP). *HPB (Oxford)* 2009;**11**:565–9.
29. Noyez L. Control charts, cusum techniques and funnel plots. a review of methods for monitoring performance in healthcare. *Interact Cardiovasc Thorac Surg* 2009;**9**:494–9.
30. Bianchi S, Martinoli C. Shoulder: normal ultrasound findings and examinationning technique. In: Baert AL, Knauth M, Sartor K, editors. *Ultrasound of the musculoskeletal system*. Berlin: Springer; 2007. p. 216–23.