BJO

HIP

Assessment of technical skill in hip fracture surgery using the postoperative radiograph

PILOT DEVELOPMENT AND VALIDATION OF A FINAL PRODUCT ANALYSIS CORE OUTCOME SET

H. K. James,
G. T. R. Pattison,
J. Griffin,
J. D. Fisher,
D. R. Griffin

From University of Warwick, Coventry, UK

Aims

To develop a core outcome set of measurements from postoperative radiographs that can be used to assess technical skill in performing dynamic hip screw (DHS) and hemiarthroplasty, and to validate these against Van der Vleuten's criteria for effective assessment.

Methods

A Delphi exercise was undertaken at a regional major trauma centre to identify candidate measurement items. The feasibility of taking these measurements was tested by two of the authors (HKJ, GTRP). Validity and reliability were examined using the radiographs of operations performed by orthopaedic resident participants (n = 28) of a multicentre randomized controlled educational trial (ISRCTN20431944). Trainees were divided into novice and intermediate groups, defined as having performed < ten or ≥ ten cases each for DHS and hemiarthroplasty at baseline. The procedure-based assessment (PBA) global rating score was assumed as the gold standard assessment for the purposes of concurrent validity. Intra- and inter-rater reliability testing were performed on a random subset of 25 cases.

Results

In total, 327 DHS and 248 hemiarthroplasty procedures were performed by 28 postgraduate year (PGY) 3 to 5 orthopaedic trainees during the 2014 to 2015 surgical training year at nine NHS hospitals in the West Midlands, UK. Overall, 109 PBAs were completed for DHS and 80 for hemiarthroplasty. Expert consensus identified four 'final product analysis' (FPA) radiological parameters of technical success for DHS: tip-apex distance (TAD); lag screw position in the femoral head; flushness of the plate against the lateral femoral cortex; and eight-cortex hold of the plate screws. Three parameters were identified for hemiarthroplasty: leg length discrepancy; femoral stem alignment; and femoral offset. Face validity, content validity, and feasibility were excellent. For all measurements, performance was better in the intermediate compared with the novice group, and this was statistically significant for TAD ($p < 0.001$) and femoral stem alignment ($p = 0.023$). Concurrent validity was poor when measured against global PBA score. This may be explained by the fact that they are measuring difference facets of competence. Intra-and inter-rater reliability were excellent for TAD, moderate for lag screw position (DHS), and moderate for leg length discrepancy (hemiarthroplasty). Use of a large multicentre dataset suggests good generalizability of the results to other settings. Assessment using FPA was time- and cost-effective compared with PBA.

Conclusion

Final product analysis using post-implantation radiographs to measure technical skill in hip fracture surgery is feasible, valid, reliable, and cost-effective. It can complement traditional workplace-based assessment for measuring performance in the real-world operating room . It may have particular utility in competency-based training frameworks and for assessing skill transfer from the simulated to live operating theatre.

Correspondence should be sent to Hannah K James; email: h.smith.1@warwick.ac.uk

doi: 10.1302/2633-1462.19.BJO-2020-0101.R1

Bone Joint Open 2020;1-9:594–604.

Cite this article: *Bone Joint Open* 2020;1-9:594–604.

Keywords: Surgical education, Technical skills, Assessment, Outcome validation

Introduction

Post-implantation radiographs are routinely used in orthopaedic practice to assess the success of hip fracture surgery and to predict risk of fixation failure. The position of the implant is widely believed to be an important factor in predicting clinical outcome[1-4] and has been repeatedly shown in the simulation laboratory to be influenced by the technical skill of the surgeon.[5-10]

With the notable exception of tip-apex distance (TAD) for dynamic hip screw (DHS),[4] there is a paucity of published evidence on the relationships between postoperative implant position, patient outcome following hip fracture surgery, and technical skill of the surgeon in the real-world clinical environment. In the absence of accepted criteria, judgement as to the satisfactory position of the implant in DHS and hemiarthroplasty appear to be made in everyday clinical practice using a global, qualitative 'expert eye' judgement, refined through experience.

The need to define a core radiological outcome set to assess a technically successful hip fracture operation is driven by the requirement in both the surgical training and educational research settings for a technical skills outcome measure that is clinically relevant and objectively measurable, reproducible, and reliable. The use of patient-centred outcome measures is key to being able to demonstrate the highest level of evidence of learning according to Kirkpatrick's hierarchy (level 4; patient results).[11] This is important for two reasons. First, in an increasingly competency-based training climate, residents must objectively demonstrate attainment of surgical skill in core procedures.[12] Second, demonstrating skills transfer to the operating theatre with resultant patient benefit is necessary to justify financial investment decisions around simulation provision.

Measurement of technical skill in the real-world orthopaedic theatre is fraught with methodological challenges, and a recent systematic review showed that none of the technical skills assessment tools in current use in orthopaedic training around the world satisfy the Norcini criteria for effective assessment.[13] Most have not been validated beyond the simulated environment, and are unsuitable for use in real life due to reliance on simulator-derived metrics for assessment of technical skill.

There is growing interest in the use of 'final product analysis' (FPA) to objectively assess the real-world technical skill of the trainee orthopaedic surgeon. There is an increasing body of evidence to show that FPA in the simulation laboratory is face,[14] content,[5,14] construct,[5-7,9,14-17] and concurrent[14] valid, and educationally impactful[5,18] for orthopaedic surgery. There is no evidence to date of the utility of FPA in the clinical setting using real patient operations.

Postoperative radiographs are an attractive candidate for FPA in the real-world clinical setting as they are objective, proximate to the time of surgery, non-invasive, and routinely collected as part of usual care. They are a useful surrogate measure where measurement of traditional gold-standard clinical outcomes such as revision rate and mortality is impractical. The ideal radiological measures for this purpose are those that are easily perceptible on a radiograph, that are clinically relevant, and which have sufficiently high resolution to be responsive to small incremental changes in technical skill.

This study is the first investigation into the real-world utility of using postoperative patient radiographs for technical skills assessment of junior surgeons. Our hypothesis is that postoperative patient radiographs can be used to measure technical skill in junior residents performing hip fracture surgery on real patients in the operating theatre, and that this will satisfy four of the domains of effective assessment described by Van der Vleuten:[19] validity; reliability; feasibility; and cost-effectiveness.

Methods

National research ethics approval was granted for this study by the NHS Research Authority South Birmingham Research Ethics Committee (15/WM/0464). Confidentiality Advisory Group approval was granted for accessing radiological data without patient consent (16/CAG/0125).

Phase 1: Consensus exercise to define core outcomes. An informal scoping literature review was undertaken to identify current evidence for assessing technical skill using postoperative radiographs in hip fracture surgery. When none was found, the focus of the scoping review was moved to look for evidence of radiological features of DHS and hemiarthroplasty that predict clinical outcome. There were no studies found relating hemiarthoplasty implant position to clinical outcome, and so the total hip arthroplasty literature was used. A list of candidate measurements was developed from the scoping literature search and externally checked with internationally recognized experts in the field to ensure that they were in line with leading opinion.

An e-Delphi exercise was undertaken to systematically combine expert opinion and achieve consensus where none currently exists. Consensus was determined to have been reached when there was ≥ 75% panel agreement, which is a widely accepted benchmark in the consensus-setting literature.[20] The consultant orthopaedic surgeon cohort in a major regional trauma centre in the UK were invited to participate (n = 39). Nineteen consultants completed all three survey rounds (49%). All consultant orthopaedic surgeons were invited regardless of subspecialism or involvement with the on-call trauma service, as hip fracture operations are a basic core trauma procedure

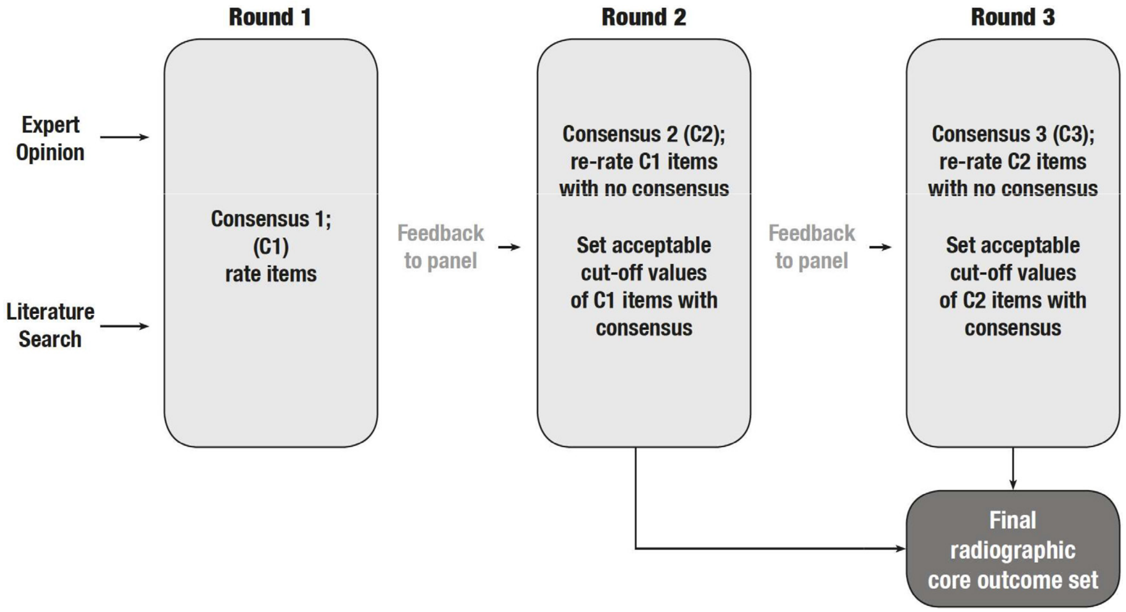

Round 1
Round 2
Round 3
Expert Opinion
Literature Search
Consensus 1; (C1) rate items
Feedback to panel
Consensus 2 (C2); re-rate C1 items with no consensus
Set acceptable cut-off values of C1 items with consensus
Feedback to panel
Consensus 3 (C3); re-rate C2 items with no consensus
Set acceptable cut-off values of C2 items with consensus
Final radiographic core outcome set

Fig. 1

Overview of the consensus process.

in which independent competence is required before completion of surgical training.[21] The Delphi panel demographic information is shown in Supplementary Table i.

The survey was built using an online survey platform (Survey Monkey Inc, San Mateo, California, USA) and administered in three rounds. An overview of the Delphi process is shown in Figure 1. In round 1, participants were presented with the candidate measurements and given binary yes/no answer options to indicate if they believed each of the proposed measures were important for assessing the skill of DHS or hemiarthroplasty. There was free text space to record opinion. In round 2, items that had achieved consensus were re-presented with proposed cut-off thresholds of acceptability, in a binary yes/no format, and free text space was provided for elaboration. Items that had not reached consensus were re-presented with the level of participant agreement (expressed as percentage) with additional details of supporting literature evidence. In round 3, items that had still not achieved consensus were represented with new acceptability threshold proposals in line with panel opinion from round 2, along with relevant supporting published evidence where appropriate. Items that failed to reach consensus after three rounds were abandoned.

Phase 2: Feasibility testing. The feasibility of obtaining the measurements identified in phase 1 was assessed by two of the authors (HKJ and GTRP). Measurements were taken within the hospital electronic Picture Archiving and Communication System (PACS) using the inbuilt user interface tools. Intraoperative Image Intensification (II) images were used for DHS, as postoperative radiographs for DHS are not routinely taken in UK orthopaedic practice. The II pictures are not autocalibrated within PACS and so were manually scaled using a known fixed implant dimension (the outer thread diameter of the DHS lag screw).

Phase 3: Validity and reliability testing. Radiographs of operations performed by orthopaedic trainee participants of a multicentre randomized controlled educational trial[22] (ISRCTN20431944) were used for validity and reliability testing. Cases were identified from the electronic surgical logbooks of operations performed by trial participants. The corresponding radiographs were retrieved from the hospital servers.

Face and content validity were addressed in Phase 1. Construct validity, the ability of an assessment instrument to discriminate between experience levels, was measured by novice and intermediate-level trainee performance over the same time period and setting(s). Novice trainees were defined as having performed < ten DHS or hemiarthoplasty cases at baseline, and ‘intermediate’, defined as having performed ≥ ten DHS or hemiarthroplasty cases at baseline. Classification of trainee experience was independently assessed for each procedure. Ten cases was chosen because previous learning curve analysis of trainees performing simulated hip fracture osteosynthesis suggests that around ten repetitions are required for performance to stabilize in the associative learning phase.[23] For continuous outcomes, we compared the means between both groups using t-test, and tested whether the difference between groups was zero. For categorical outcomes, we conducted a chi-squared test of association, or Fisher’s exact test if cell counts were less than five.

Concurrent validity, the performance of an assessment instrument against the current gold standard, was determined by comparing performance as measured by implant position on the radiographs against the global rating scale component of the procedure-based assessment (PBA) scores. The PBA is the current gold standard summative assessment tool used in higher orthopaedic surgical training in the UK.[24] They are collected routinely during training, although not mandated for every case. We assessed the same outcome measures for both procedures described above.

All primary measurements were taken by one author (HKJ, orthopaedic trainee). An adequate reliability testing sample size was determined to be 25 cases. A randomly selected subset of 25 DHS and 25 hemiarthroplasty cases were re-measured on two occasions one week apart to determine intra-rater reliability, and on one occasion by a second rater (GTRP, attending orthopaedic surgeon) to determine inter-rater reliability.

We conducted both intra-and inter-rater reliability analyses to assess the reliability of the primary rater, and the comparability of measures with the independent rater, respectively. For measures which were continuous, we plotted Bland-Altman plots to assess differences in measures and then calculated intraclass correlation coefficients to describe how strongly associated the scores were with accompanying 95% confidence intervals. For categorical outcomes, we used an equivalent measure for assessing agreement, the Cohen’s kappa statistic, and the crude percentage agreement in absolute terms.

Table I. Baseline characteristics of surgeons.*

DHS	Novice (n = 13)	Intermediate (n = 15)
Mean age in years (SD)	28.8 (4.63)	29.9 (3.28)
Range	25 to 42	26 to 37
Male	5 (38%)	3 (20%)
Female	8 (62%)	12 (80%)
Mean completed months T&O training at baseline (SD)	6.42 (4.89)	20.73 (13.9)
Range	0 to 15	2 to 54
Mean DHS cases performed at baseline (SD)	3.08 (3.40)	19.20 (6.50)
Range	0 to 9	11 to 33
Hemiarthroplasty	**Novice (n = 19)**	**Intermediate (n = 9)**
Mean age in years (SD)	28.6 (3.86)	31.22 (3.63)
Range	25 to 42	26 to 37
Male	14 (74%)	6 (67%)
Female	5 (26%)	3 (33%)
Mean completed months T&O training at baseline (SD)	8.06 (5.46)	27.00 (14.53)
Range	0 to 20	2 to 54
Mean hemiarthroplasty cases performed at baseline (SD)	2.05 (2.78)	21.11 (5.82)
Range	0 to 8	13 to 29

*Reported by operation type as some residents were in the novice group for one procedure and intermediate group for the other.
DHS, dynamic hip screw; SD, standard deviation; T&O, trauma and orthopaedics.

Results

Overall, 28 core trainee 1 (CT1) to specialty trainee 3 (ST3) trainees performed 327 DHS operations and 248 hemiarthroplasty operations during one surgical training year (August 2014 to August 2015) in nine regional NHS hospitals in the UK. There were 109 PBAs completed for DHS and 80 for hemiarthroplasty in the study population. Baseline demographics of the trainee participants are shown in Table I. Only operations coded as ‘supervised trainer scrubbed’ or ‘supervised trainer unscrubbed’ were included in the analysis, to ensure that the included operations were actually performed by the trainee participants. Operations coded as ‘performed’ were excluded as these are unsupervised, non-training operations and therefore there would not be a corresponding PBA completed.

Face and content validity. Face validity (that a tool is fit for purpose) and content validity (that a tool tests appropriate domains) can both be demonstrated through expert consensus-setting exercises. Candidate items were externally checked by recognized international experts in hip fracture surgery. The items that achieved consensus > 75% through the e-Delphi process, with descriptors and acceptability thresholds, are shown in Table II.

Four FPA radiological parameters were identified for DHS: tip-apex distance; lag screw position in the femoral head with reference to Cleveland’s zones;[25] flushness of the plate against the lateral femoral cortex; and eight-cortex hold of the plate screws (Figures 2 and 3). Three radiological parameters were identified for hemiarthroplasty: leg-length discrepancy; femoral stem alignment; and femoral offset (Figure 4). Rejected items were ‘cortical screws perpendicular to plate’ for DHS, which was rejected 68% against in round one, and ‘cement thickness’ for hemiarthroplasty, which failed to reach consensus after three rounds.

A schematic diagram of the measurement parameters is shown in Figures 2–4.

Construct validity. Construct validity (the discriminant ability of a test instrument to distinguish between experience levels) was evaluated by comparing between-group differences for the various metrics. Results of construct validity testing are shown in Table III (DHS) and IV (hemiarthroplasty). For DHS, TAD as a continuous variable was found to be significantly different between experience levels, with the intermediate group having a lower mean TAD, signifying a technically superior result, 18.3 mm compared with 15.7 mm for novices and intermediates, respectively, $p < 0.001$.

Table II. Candidate item inclusion and exclusion by Delphi round.

Dynamic hip screw		
Included items		
Item	**Round in which consensus was achieved**	**% agreement**
1a. Tip-apex distance (TAD)	1	88
1b. Acceptable TAD < 25 mm	2	90
2a. Lag screw position in femoral head	2	90
2b. Described according to Cleveland's 9 zones	2	90
3a. Plate position	1	83
3b. Acceptable = plate flush with cortex on AP, no gaps seen	1	83
4a. Cortical screw position	1	88
4b. Acceptable = 8 cortex hold	1	88
Excluded Items		
Item	Exclusion reason	
Screws perpendicular with plate	Rejected, 68% against in round 1	
Hemiarthroplasty		
Leg length discrepancy (LLD)	1	92
Acceptable LLD = ≤ 15 mm	3	89
Femoral stem alignment (FSA)	1	88
Acceptable alignment = ≤ or ≥ 5 ° from neutral	3	95
Femoral offset	2	81
Acceptable = should be equal to native side	2	81
1. Excluded Items		
Item	Exclusion reason	
Cement thickness	Failed to reach consensus after 3 rounds.	

AP, anteroposterior.

Tip-apex distance < 25 mm as a dichotomous variable was not discriminant between the two groups (p = 0.222).

Mean PBA scores for DHS were seen to improve significantly between the novice group with a mean global rating score of 2.4 and the intermediate group with a mean score of 2.8 (p < 0.001). There was no difference seen in lag screw position in the femoral head between the two groups (p = 0.393). There were fewer plates flush to the lateral femoral cortex in the novice group (58%) as compared with the intermediate group (66%), but this was not statistically significant (p = 0.153). Similarly, there were slightly more procedures that failed to demonstrate eight-cortex hold in the novice group (4%) compared with the intermediate group (1%) but this difference was again not significant.

For hemiarthroplasty, femoral stem alignment was found to be significantly better in the intermediate group than in the novice group, with a mean deviation from neutral of 3.1° for novices and 2.6° for intermediates; p = 0.023. Leg length discrepancy and femoral offset difference were both found to be better in the intermediate group as compared with the novices, but these differences were not statistically significant.

The intermediate group achieved a significantly higher mean global PBA score than the novice group for hemiarthroplasty; the mean score was 2.4 for novices and 2.9 for intermediates; p < 0.001.

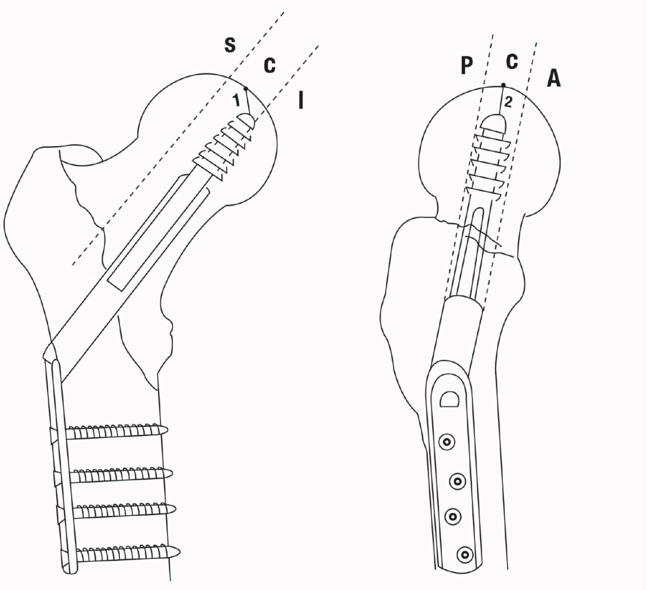
Concurrent validity. Concurrent validity was measured by examining differences between global PBA scores (treated as categorical variables) and each of the radiological measurements for DHS and hemiarthroplasty (Table IV to Table V). No significant association between PBA global rating score and any of the seven tested radiological parameters was found using the chi-squared test.

Reliability. For DHS, both intra- and inter-rater reliability were found to be excellent for TAD (Cohen's kappa 0.84 and 0.76), and moderate for position of lag screw (Cohen's kappa 0.47 for both intra- and inter-rater reliability) (Table VI). Intra-rater reliability was found to be poor for assessing whether or not the plate was flush to the lateral cortex of the femur (Cohen's kappa 0.12). The Kappa statistic could not be calculated for intra- and inter-rater reliability for eight-cortex hold and for inter-rater reliability for plate flush to femur due to one rater having no variation in measurement.

For hemiarthroplasty, the intra-rater reliability was found to be moderate for leg length discrepancy and femoral stem alignment (Cohen's kappa 0.57 and 0.59, respectively), and excellent for femoral offset difference (Cohen's kappa 0.79). The inter-rater reliability was moderate for leg length discrepancy (Cohen's kappa 0.54), fair for femoral stem alignment (Cohen's kappa 0.33), and poor for femoral offset difference (Cohen's kappa 0.18)(Table VII).

Cost-effectiveness. The radiographs that were measured were collected as a routine part of intra/postoperative care, and therefore represented no extra cost burden from an educational assessment or clinical care point of view. Assessor time in taking the measurements was recorded as a mean of 45 seconds per case for DHS and 57 seconds per case for hemiarthroplasty. This is significantly lower than the recommended average time to complete a PBA form of ten to 15 minutes.[24]

To be sure that the case mix encountered by the two groups was comparable, we classified the hip fractures into simple/moderate/complex for DHS and simple/complex for hemiarthroplasty, based around the AO classification system.[26] We found no differences in the fracture complexity between the novice and intermediate groups for either DHS or hemiarthroplasty (classification matrix, and table in Supplementary Tables ii and iii).



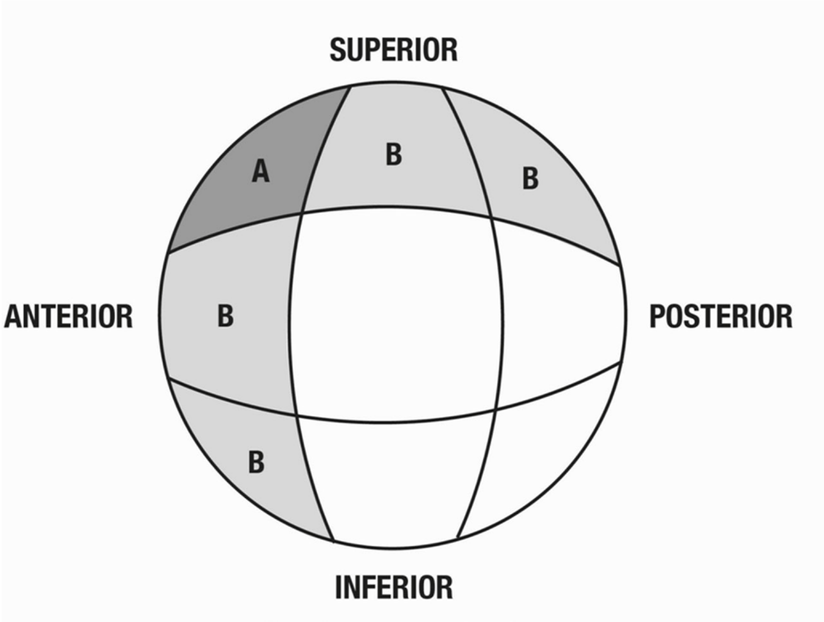
Tip Apex Distance (TAD) = Distance 1(mm) + Distance 2(mm)

S = Superior
C = Central
I = Inferior

P = Posterior
C = Central
A = Anterior

Fig. 2

Schematic diagram of dynamic hip screw measurements, anteroposterior and lateral views.



Zone A = anterior **and** superior position

Zone B = anterior **or** superior position

Fig. 3

View looking at femoral head showing modified Cleveland zones.

Discussion

The ultimate goal of surgical training is to produce safe surgeons who perform good quality operations for their patients. The use of FPA to assess surgical skill using patient radiographs may help bridge the perceived gap between educational assessment and real-world clinical performance. Postoperative radiographs are a promising resource for real-world FPA as part of the move towards competency assessment in training, and also to measure transfer of skills from the simulated environment.

This is the first study to explore the use of patient radiographs for FPA assessment of technical skill, and we have systematically addressed the key domains of effective assessment: face, content, construct, and concurrent validity; feasibility; reliability; and cost-effectiveness.

Our results showed reasonable face and content validity of the radiological outcome measures within the limits of a Delphi exercise. As we are assessing the role of radiological FPA as a surrogate for clinical outcome in an educational assessment setting, it is difficult to show

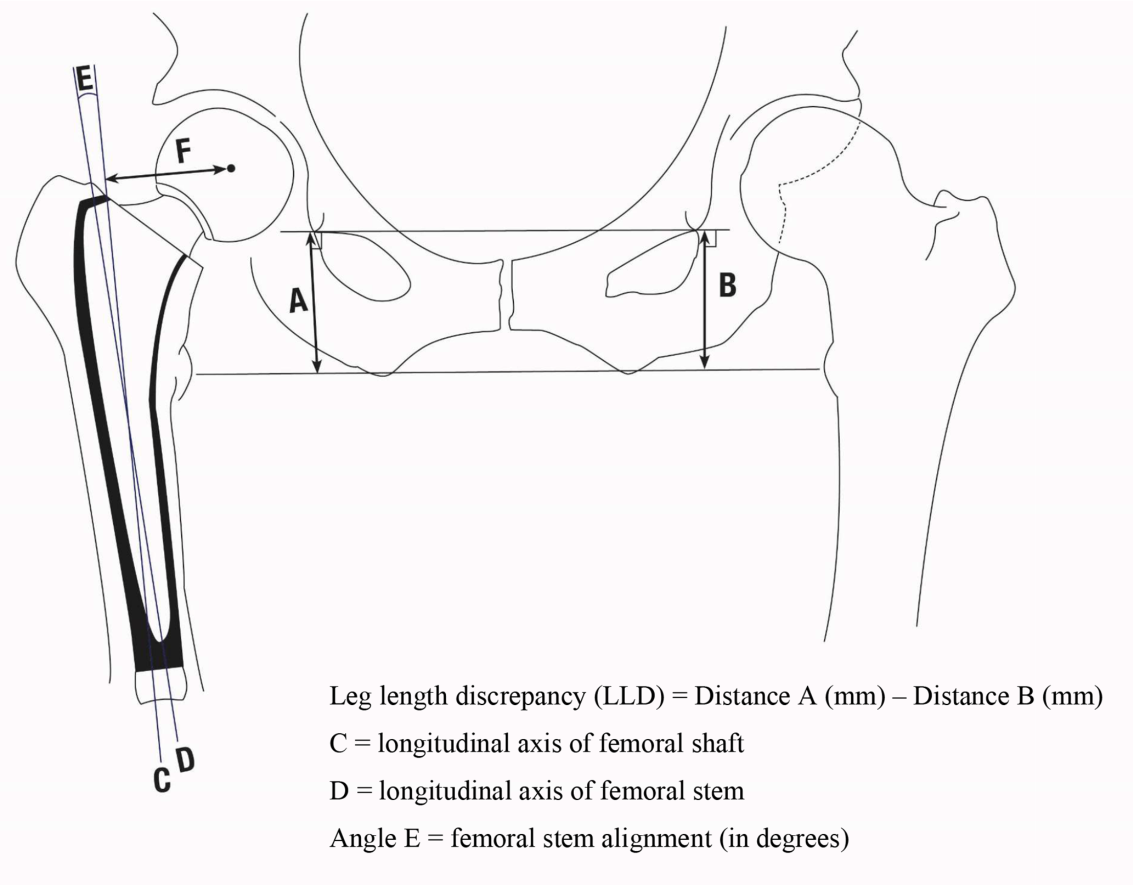

E
F
A
B
C
D
Leg length discrepancy (LLD) = Distance A (mm) – Distance B (mm)
C = longitudinal axis of femoral shaft
D = longitudinal axis of femoral stem
Angle E = femoral stem alignment (in degrees)

Fig. 4

Schematic diagram of hemiarthroplasty measurements, anteroposterior view.

comprehensiveness and comprehensibility as is traditionally required of content validity. It is reasonable to say that this outcome set appears to be relevant for the population (trainee surgeons) and context of interest (hip fracture surgery).

The construct validity picture was mixed, with one of four measures for DHS (TAD) and one of three measures for hemiarthroplasty (femoral stem alignment) demonstrating statistically significant differences between groups divided by experience level. The trend across all measurements showed improvement between the novice and intermediate groups, suggesting evidence of construct validity, although we cannot claim construct validity of the tool as a whole based on this pilot study. The construct validity of the global rating scale for PBA was found to be excellent for both DHS and hemiarthroplasty.

The concurrent validity with PBA was explored by comparing the radiological parameters for operations which scored a PBA level 2 or 3 (Table VIIIa IVb). We did not show evidence of concurrent validity to our gold standard. The use of PBA as gold standard, as opposed to a clinical outcome, is a significant limitation of this study. None of the parameters demonstrated a statistically significant relationship with PBA global rating scale score. This is the first investigation of the association between PBA and the quality of the outcome of the operation as measured by the radiograph. This finding might be a reflection of the fact that they are assessing different things; the PBA global rating scale is designed to assess the overall ability of the trainee to perform the procedure without supervision, rather than to assess the quality of the operation or technical skill in doing so. Hence the PBA and our radiological outcome measurements are assessing two different facets of competence that are not directly comparable, which may explain the apparent observed lack of concurrent validity.

The intra-rater reliability was generally excellent for both DHS (with the exception of plate flush to femur) and hemiarthroplasty, and the inter-rater reliability was generally excellent for DHS, but moderate to poor for hemiarthroplasty. This finding might be explained by the fact that the measurement technique is more readily standardized for DHS, whereas there is greater scope for subjectivity in deciding on appropriate landmarks for measuring leg length discrepancy and offset. This is likely to be compounded by the fact that the postoperative films were often of poor quality, supine and rotated, in contrast to standing films seen in the elective arthroplasty setting.

Table III. Construct validity of dynamic hip screw radiological outcome measures.

Outcome	Variable	Novice (residents < 10 cases)	Intermediate (residents ≥ 10 cases)	Total	p-value
Tip-apex distance (TAD), mm	Mean (SD)	18.3 (6.4)	15.7 (5.8)	16.8 (6.2)	< 0.001
	Number	138	185	323	
	Range	5.9 to 46.7	6.2 to 38.7	5.9 to 46.7	
Position of lag screw in femoral head	Neither superior nor anterior	109 (76%)	137 (76%)	246 (76%)	0.393
	Superior or anterior	34 (24%)	41 (23%)	75 (23%)	
	Superior and anterior	0 (0%)	3 (1%)	3 (1%)	
	Total	143	181	324	
Plate flush to femur	Yes	83 (58%)	121 (66%)	204 (62%)	0.153
	No	60 (42%)	63 (34%)	123 (38%)	
	Total	143	184	327	
Eight-cortex hold	Yes	135 (96%)	181 (99%)	316 (98%)	0.082
	No	6 (4%)	2 (1%)	8 (2%)	
	Total	141	183	324	
PBA global rating scale (continuous)	Mean (SD)	2.4 (0.6)	2.8 (0.5)	2.6 (0.6)	<0.001
	Number	58	60	118	
	Range	1 to 4	2 to 3	1 to 4	
PBA global rating scale (categorical)	1	1 (2%)	0 (0%)	1 (1%)	0.001
	2	34 (59%)	16 (27%)	50 (42%)	
	3	22 (38%)	42 (70%)	64 (54%)	
	4	1 (2%)	2 (3%)	3 (3%)	
	Total	58	60	118	

*Continuous variables: *t*-test comparing means, Categorical variables: chi-squared test where cells > five cases, Fisher's exact test where cells ≤ five cases.
PBA, procedure-based assessment.

Table IV. Concurrent validity of dynamic hip screw radiological outcome measures.

Outcome	Variable	Novice (residents < 10 cases)	Intermediate (residents ≥ 10 cases)	p-value
Leg length discrepancy, mm	Mean (SD)	3.7 (3.2)	3.5 (2.7)	0.756
	Number	40	32	
	Range	0.1 to 13.0	0.1 to 10.2	
Femoral stem alignment, degrees	Mean (SD)	2.4 (1.4)	2.7 (2.0)	0.442
	Number	37	34	
	Range	0.2 to 7.6	0.1 to 7.7	
Femoral offset difference, mm	Mean (SD)	9.9 (5.9)	10.3 (8.7)	0.792
	Number	39	34	
	Range	0.1 to 24.5	0.1 to 43.6	

*Continuous variables: *t*-test comparing means, Categorical variables: chi-squared test where cells > five cases, Fisher's exact test where cells ≤ five cases.

The feasibility was excellent, with the measurements easy and quick to obtain using readily accessible technology. The cost-effectiveness was also superficially excellent, with no additional cost associated with the radiographs other than the assessors' time. Time-to-assess per case was substantially lower for FPA using postoperative radiographs than for PBA completion by an order of magnitude of at least ten-fold.

A strength of our study is that we have systematically assessed nearly 600 real operations performed by 28 trainees across nine hospital sites over one surgical training year, which is a much larger sample with longer follow-up than most educational studies. The large sample size and the multicentre nature of the data suggest that the generalizability of our results is good and the chance of a type 2 error small.

This study has several weaknesses. Our scoping review was informal, and therefore it is possible some outcomes could have been missed. We only considered four of five of the Van der Vleuten's utility domains of effective assessment, as we have excluded 'educational impact'. This decision was taken because separate, qualitative assessment of the educational impact of using radiological measurements for learning would be required and this analysis was conducted retrospectively. Previous work has shown that the morning trauma meeting, where radiographs are displayed and discussed, is educationally valuable for trainees.[27] Other simulation-based studies have shown

Table V. Concurrent validity of dynamic hip screw radiological outcome measures.

Outcome	Variable	PBA score 2	PBA score 3	p-value
Tip-apex distance (TAD), mm	Mean (SD)	17.2 (6.4)	17.5 (6.5)	0.815
	Number	51	63	
	Range	5.9 to 39.3	8.4 to 38.0	
Position of lag screw in femoral head, n (%)	Neither superior nor anterior	37 (74)	46 (73)	0.439
	Superior or anterior	13 (26)	15 (24)	
	Superior and anterior	0 (0)	2 (3)	
	Total	50	63	
Plate flush to femur, n (%)	Yes	29 (57)	40 (63)	0.540
	No	22 (43)	24 (37)	
	Total	51	64	
Eight-cortex hold, n (%)	Yes	47 (94)	64 (100)	0.082
	No	3 (6)	0 (0)	
	Total	50	64	

*Continuous variables: *t*-test comparing means, Categorical variables: chi-squared test where cells > five cases, Fisher's exact test where cells ≤ five cases.
PBA, procedure-based assessment.

Table VI. Intra- and inter-rater reliability of dynamic hip screw outcome measures.

Outcome	Measure	Intra-rater (Rater 1 vs Rater 1)	Inter-rater (Rater 1 vs Rater 2)
Tip-apex distance (TAD), mm	ICC (95% CI)	0.835 (0.66 to 0.92)	0.763 (0.53 to 0.88)
	Mean difference (95% LOI)	2.88 (-1.37 to 7.13)	2.24 (-3.14 to 7.16)
Position of lag screw in femoral head	Cohen's κ (95% CI)	0.468 (-0.18 to 1.00)	0.468 (-0.18 to 1.00)
	Percentage agreement (95% CI)	92% (81% to 100%)	92% (81% to 100%)
Plate flush to femur	Cohen's κ (95% CI)	0.123 (-0.03 to 0.27)	*
	Percentage agreement (95% CI)	43% (22% to 65%)	33% (13% to 54%)
Eight-cortex hold	Cohen's κ (95% CI)	*	*
	Percentage agreement (95% CI)	91% (79% to 100%)	90% (78% to 100%)

*Kappa statistic not calculated due to one rater having no variation so kappa statistic cannot be calculated.
CI, confidence interval.; k, kappa.; LOI, limits of agreement.

Table VII. Intra- and inter-rater reliability of hemiarthroplasty outcome measures.

Outcome	Measure	Intra-rater (Rater 1 vs Rater 1)	Inter-rater (Rater 1 vs Rater 2)
Leg length discrepancy	ICC (95% CI)	0.573 (0.23 to 0.79)	0.541 (0.18 to 0.77)
	Mean difference (95% LOI)	0.22 (-0.53 to 0.96)	0.03 (-0.46 to 0.53)
Femoral stem alignment	ICC (95% CI)	0.594 (0.26 to 0.80)	0.326 (0.08 to 0.64)
	Mean difference (95% LOI)	0.57 (-3.48 to 4.61)	0.41 (-2.68 to 3.51)
Femoral offset difference	ICC (95% CI)	0.790 (0.57 to 0.90)	0.18 (0,0.54)
	Mean difference (95% LOI)	0.40 (-3.32 to 2.52)	0.05 (-0.82 to 0.93)

CI, confidence interval; ICC, intraclass correlation coefficient; LOI, limits of agreement.

FPA to be an educationally valuable assessment method in orthopaedic surgery.[5,18] It is therefore not unreasonable to assume that FPA, with appropriately delivered feedback, would be educationally impactful, although we did not specifically seek to address this in our study. Our gold standard, the PBA, may not have been the best comparator. Ideally, we would have compared the radiological outcomes with clinical outcomes.

With the probable exception of TAD, given the weight of evidence supporting its clinical relevance and clearly significant construct validity demonstrated in our results, the measurements we have defined here are unlikely to be useful in isolation for assessing competence. Rather, they may be most useful as an adjunct to traditional technical skills assessment in the workplace, to help overcome the well-recognized limitations of these. The pilot FPA outcome sets we have described here may also be useful in developing competency thresholds for simulation-based training. It is possible that the radiological metrics we have investigated in this study could be combined into a composite score, and further work is needed to ascertain appropriate weightings for the individual items and to pilot test these.

Table VIII. Construct validity of hemiarthoplasty radiological outcome measures.

Outcome	Variable	Novice (residents < 10 cases)	Intermediate (residents ≥ 10 cases)	Total	p-value*
Leg length discrepancy, mm	Mean (SD)	4.1 (3.4)	3.6 (3.2)	3.9 (3.3)	0.233
	Number	120	119	239	
	Range	0 to 13.6	0 to 17.5	0 to 17.5	
Femoral stem alignment (degrees)	Mean (SD)	3.1 (2.0)	2.6 (1.4)	2.9 (1.7)	0.023
	Number	120	115	235	
	Range	0.1 to 9.6	0.1 to 7.6	0.1 to 9.6	
Femoral offset difference, mm	Mean (SD)	9.0 (7.5)	7.9 (6.8)	8.5 (7.2)	0.246
	Number	123	125	248	
	Range (SD)	0 to 43.6	0 to 44.1	0 to 44.1	
PBA Global rating scale (continuous)	Mean (SD)	2.4 (0.5)	2.9 (0.6)	2.6 (0.6)	< 0.001
	Number	47	29	76	
	Range	2 to 3	2 to 4	2 to 4	
PBA Global rating scale (categorical), n (%)	1	0 (0)	0 (0)	0 (0)	0.002
	2	29 (62)	8 (28)	37 (49)	
	3	18 (38)	17 (59)	35 (46)	
	4	0 (0)	4 (13)	4 (5)	
	Total	47	29	76	

*Continuous variables: *t*-test comparing means, Categorical variables: chi-squared test where cells > five cases, Fisher's exact test where cells ≤ five cases.
PBA, procedure-based assessment; SD, standard deviation.

Conclusion

It is feasible to measure technical skill in orthopaedic trainees performing hip fracture surgery using intra- or postoperative patient radiographs, and this is probably cost-effective, and appears to be face and content valid. Performance was widely observed to be better in the intermediate than in the novice group suggestive of construct validity, and this was statistically significant for TAD in DHS and femoral stem alignment in hemiarthroplasty. Improvement in these measures with increased experience suggest that they are responsive to small incremental changes in technical skill. Concurrent validity was poor when measured against the PBA global rating scale score, but this may be because the PBA is not designed to assess technical skill. Procedure-based assessment may not be the best gold standard measure. Intra- and inter-rater reliability were variable, and found to be excellent for TAD, and moderate for lag screw position (DHS) and leg length discrepancy (hemiarthroplasty). Use of a large, longitudinal, multicentre educational trial dataset suggests the generalizability of these results is good. The FPA using patient radiographs is likely to be most useful as part of a battery of assessment of technical skill, and may have a role in complementing traditional workplace-based assessment in determining technical skill in the real-world OR. It may have particular utility in competency-based training frameworks, and for assessing skill transfer from the simulated to live operating theatre. These results should be regarded as provisional, and until further validation evidence is provided, the PBA remains the best current tool for assessing technical skill in surgical trainees.

Take home message

- Post/intra-operative radiographs can be used to assess technical skill in hip fracture surgery. This can complement traditional workplace-based assessment for measuring operative performance and may have particular value in competency-based surgical training.

Twitter

Follow H. K. James @hannah_ortho
Follow G. T. R. Pattison @pattison_giles
Follow D. R. Griffin @DamianGriffin

Supplementary material

Tables showing demographics of Delphi Panel, fracture complexity by surgeon experience level, and fracture complexity codes by AO classification.

References

1. **Srivastav S**, **Mittal V**, **Agarwal S**. Total hip arthroplasty following failed fixation of proximal hip fractures. *Indian J Orthop*. 2008;42(3):279–286.
2. **Kammerlander C**, **Neuerburg C**, **Gosch M**, **Böcker W**. Patient outcomes after screw fixation of hip fractures. *Lancet*. 2018;392(10161):2264–2265.
3. **Parker MJ**, **Kendrew J**, **Gurusamy K**. Radiological predictive factors in the healing of displaced intracapsular hip fractures. A clinical study of 404 cases. *Hip Int*. 2011;21(4):393–398.
4. **Baumgaertner MR**, **Curtin SL**, **Lindskog DM**, **Keggi JM**. The value of the tip-apex distance in predicting failure of fixation of peritrochanteric fractures of the hip. *J Bone Joint Surg Am*. 1995;77-A(7):1058–1064.
5. **Christian MW**, **Griffith C**, **Schoonover C**, **et al**. Construct validation of a novel hip fracture fixation surgical simulator. *J Am Acad Orthop Surg*. 2018;26(19):689–697.
6. **Sugand K**, **Wescott RA**, **Carrington R**, **Hart A**, **Van Duren BH**. Teaching basic trauma: validating FluoroSim, a digital fluoroscopic simulator for guide-wire insertion in hip surgery. *Acta Orthop*. 2018;89(4):380–385.
7. **Akhtar K**, **Sugand K**, **Sperrin M**, **et al**. Training safer orthopedic surgeons. Construct validation of a virtual-reality simulator for hip fracture surgery. *Acta Orthop*. 2015;86(5):616–621.

8. **Froelich JM**, **Milbrandt JC**, **Novicoff WM**, **Saleh KJ**, **Allan DG**. Surgical simulators and hip fractures: a role in residency training? *J Surg Educ*. 2011;68(4):298–302.
9. **Nousiainen MT**, **Omoto DM**, **Zingg PO**, **et al**. Training femoral neck screw insertion skills to surgical trainees: computer-assisted surgery versus conventional fluoroscopic technique. *J Orthop Trauma*. 2013;27(2):87–92.
10. **Logishetty K**, **Rudran B**, **Cobb JP**. Virtual reality training improves trainee performance in total hip arthroplasty: a randomized controlled trial. *Bone Joint J*. 2019;101-B(12):1585–1592.
11. **Kirkpatrick D**. Techniques for evaluating training programmes. *Journal of American Society for Training and Development*. 1959;13(11):11–12.
12. **James H**. Measuring the educational impact of simulation training in Trauma & Orthopaedics. *Journal of Trauma & Orthopaedics*. 2019;7(3):54–56.
13. **James H**, **Chapman A**, **Pattison G**, **Fisher JG**, **Griffin DR**. Analysis of tools used in assessing technical skills and operative competence in trauma & orthopaedic surgical training. *JBJS(Am) Reviews*. 2020;8(6):e19.00167.
14. **Shi J**, **Hou Y**, **Lin Y**, **Chen H**, **Yuan W**. Role of Visuohaptic Surgical Training Simulator in Resident Education of Orthopedic Surgery. *World Neurosurg*. 2018;111:e98–e104.
15. **Hohn EA**, **Brooks AG**, **Leasure J**, **et al**. Development of a surgical skills curriculum for the training and assessment of manual skills in orthopedic surgical residents. *J Surg Educ*. 2015;72(1):47–52.
16. **Xiang L**, **Zhou Y**, **Wang H**, **et al**. Significance of Preoperative Planning Simulator for Junior Surgeons' Training of Pedicle Screw Insertion. *J Spinal Disord Tech*. 2015;28(1):E25–E29.
17. **Burns GT**, **King BW**, **Holmes JR**, **Irwin TA**. Evaluating internal fixation skills using surgical simulation. *J Bone Joint Surg Am*. 2017;99-A(5):e21.
18. **Bergeson RK**, **Schwend RM**, **DeLucia T**, **et al**. How accurately do novice surgeons place thoracic pedicle screws with the free hand technique? *Spine*. 2008;33(15):E50 1–E507.
19. **Van Der Vleuten CP**. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;1(1):41–67.
20. **Diamond IR**, **Grant RC**, **Feldman BM**, **et al**. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol*. 2014;67(4):401–409.
21. **Specialist Training in Trauma and Orthopaedics - 2015**. *Curriculum*. London: General Medical Council, 2018.
22. **James HK**, **Pattison GTR**, **Fisher JD**, **Griffin DR**. *Cadaveric simulation vs standard training for postgraduate trauma & orthopaedic surgical trainees: protocol for the CAD:TRAUMA study multi-centre randomised controlled educational trial*, 2020.
23. **Gustafsson A**, **Pedersen P**, **Rømer TB**, **et al**. Hip-fracture osteosynthesis training: exploring learning curves and setting proficiency standards. *Acta Orthop*. 2019;90(4):348–353.
24. **No authors listed**. COVID-19. ISCP (Intercollegiate Surgical Curriculum Programme). 2020. https://www.iscp.ac.uk (date last accessed 13 June 2020).
25. **Cleveland M**, **Bosworth DM**, **Thompson FR**, **Wilson HJ**, **Ishizuka T**. A ten-year analysis of Intertrochanteric fractures of the femur. *J Bone Joint Surg Am*. 1959;41-A(8):1399–1408.
26. **No authors listed**. COVID-19. AO Foundation. 2020. https://www.aofoundation.org/Structure/Pages/default.aspx (date last accessed 13 June 2020).
27. **Haider Z**, **Hunter A**. Orthopedic trainees' perceptions of the educational value of daily trauma meetings. *J Surg Educ*. 2020;77(4):991–998.

Author information:

- H. K. James, PhD, MRCS, Specialist Registrar, Trauma & Orthopaedic Surgery, Honorary Clinical Lecturer
- D. R. Griffin, MBM, BCh(Oxon), MA, MPhil(Cantab), FRCS, Professor of Trauma & Orthopaedic Surgery, Consultant Orthopaedic Surgeon
Clinical Trials Unit, Warwick Medical School, Coventry, UK; Department of Trauma & Orthopaedic Surgery, University Hospitals Coventry & Warwickshire, Coventry, UK.
- G. T. R. Pattison, FRCS(Tr&Orth), MMedEd, Consultant Orthopaedic Surgeon, Department of Trauma & Orthopaedic Surgery, University Hospitals Coventry & Warwickshire, Coventry, UK.
- J. Griffin, MSc, Research Associate
- J. D. Fisher, PhD, Senior Research Fellow
Clinical Trials Unit, Warwick Medical School, Coventry, UK.

Funding statement:

- This work was supported by Versus Arthritis (grant number 20485).

ICMJE COI statement:

- The authors have no financial or competing interests to disclose.

Acknowledgements:

- The authors thank Dr Helen Parsons, Associate Professor in Medical Statistics, for advising on the reliability sample size calculation, and the Trauma & Orthopaedic consultants of University Hospitals Coventry & Warwickshire for taking part in the Delphi exercise.

Ethical review statement:

- National research ethics approval was granted for this study by the NHS Research Authority South Birmingham Research Ethics Committee (15/WM/0464). Confidentiality Advisory Group approval was granted for accessing radiographic data without patient consent (16/CAG/0125).