Virtual-reality simulation to assess performance in hip fracture surgery

Poul Pedersen[1], Henrik Palm[2], Charlotte Ringsted[3], and Lars Konge[4]

[1]Department of Orthopaedic Surgery, Hospital of Nykøbing F; [2]Department of Orthopaedic Surgery, Copenhagen University Hospital, Hvidovre, Denmark; [3]Department of Anesthesia, University of Toronto, The Wilson Center, University of Toronto and University Health Network, Toronto, Canada; [4]Center for Clinical Education, University of Copenhagen and the Capital Region of Denmark, Copenhagen, Denmark.
Correspondence: pedersenpoul@dadlnet.dk
Submitted 13-09-02. Accepted 14-03-21

Background and purpose **Internal fixation of hip fractures is a common and important procedure that orthopedic surgeons must master early in their career. Virtual-reality training could improve initial skills, and a simulation-based test would make it possible to ensure basic competency of junior surgeons before they proceed to supervised practice on patients. The aim of this study was to develop a reliable and valid test with credible pass/fail standards.**

Methods **20 physicians (10 untrained novices and 10 experienced orthopedic surgeons) each performed 3 internal fixation procedures of an undisplaced femoral neck fracture: 2 hook-pins, 2 screws, and a sliding hip screw. All procedures were preformed on a trauma simulator. Performance scores for each procedure were obtained from the predefined metrics of the simulator. The inter-case reliability of the simulator metrics was explored by calculation of intra-class correlation coefficient. Validity was explored by comparison between novices' and experts' scores using independent-samples t-test. A pass/fail standard was set by the contrasting-groups method and the consequences were explored.**

Results **The percentage of maximum combined score (PM score) showed an inter-case reliability of 0.83 (95% CI: 0.65–0.93) between the 3 procedures. The mean PM score was 30% (CI: 7–53) for the novices and 76% (CI: 68–83) for the experienced surgeons. The pass/fail standard was set at 58%, resulting in none of the novices passing the test and a single experienced surgeon failing the test.**

Interpretation **The simulation-based test was reliable and valid in our setting, and the pass/fail standard could discriminate between novices and experienced surgeons. Potentially, training and testing of future junior surgeons on a virtual-reality simulator could ensure basic competency before proceeding to supervised practice on patients.**

Worldwide, hip fractures account for substantial healthcare costs and high mortality, morbidity, and reoperation rates. Fractures are often treated with different types of internal fixation, which is a great contributor to training of surgical skills through the principle of the master-apprentice model. This model is not without risk, though. Inexperienced trainees contribute to a higher rate of re-admissions and reoperations (Palm et al. 2007, Leblanc et al. 2013).

A review of 609 studies found that virtual-reality simulation training improved operative skills (Cook et al. 2011), and simulation-based training in orthopedic surgery is starting to emerge. Most of the development is in arthroscopy, but simulation-based training in fracture fixation is also being developed (Blyth et al. 2007, 2008, Mabrey et al. 2010, Atesok et al. 2012, Rambani et al. 2013). Current papers describe the simulators and explore construct validity of the simulator metrics (Tillander et al. 2004, Froelich et al. 2011). Reliable and valid tests with credible pass/fail standards are necessary to ensure basic competency of trainees before allowing them to proceed to supervised practice on patients (Stefanidis et al. 2012a, Konge et al. 2013).

The objective of this study was to develop a test, to explore the reliability of this test, and to gather validity evidence. Furthermore, we wanted to establish a credible pass/fail standard and explore the consequences of this standard. The research questions were: (1) Which simulator metrics were able to discriminate between novices and experienced orthopedic surgeons?; (2) How many procedures on the simulator must be performed to ensure sufficient reliability?; and (3) What was a credible pass/fail standard in the test?

DOI 10.3109/17453674.2014.917502

Methods

The virtual-reality simulator that we used was the Swemac TraumaVision (STV). It consists of a computer with 2 screens and TraumaVision 5.12 software. The software contains a variety of orthopedic procedures. A robot arm (Phantom Omni) is connected to a computer and mimics the operation tools, and generates haptic feedback. The robot arm can be handled by the right hand or the left hand according to the preference of the user. Fluoroscopy is administered by pressing a foot-controlled paddle, and is recorded on a standard A-P and lateral radiograph.

We developed a test by combining 3 procedures of internal fixation for hip fractures in the STV simulator. Before the test, all participants were allowed to get used to the computer simulator by placing 2 simulated distal locking screws in a femoral nail. The time limit for this "warm-up" was 20 minutes. The participants went on to perform the test as soon as the 2 distal locking screws were placed or when the "warm-up" time limit was reached. All test procedures started with placing the K-wire guide; the incision and handling of soft tissue were not simulated. The first procedure was the placement of 2 cannulated screws, and the second procedure was the insertion of 2 Hansson hook-pins. The third and final procedure was insertion of a dynamic hip screw. The 3 procedures had 5 identical simulator metrics: "Flouroscopy time", "No. of X-rays", "No. of retries in guide placement", "Procedure time in seconds", and "Score". The "Score" from the STV simulator was a combined score, and was calculated from all the procedure-specific simulator metrics. The simulator metrics that generated the "Score" was weighted depending on clinical importance and was defined by the manufacturer of the simulator. Apart from this, other simulator metrics combined in the "Score" were guide placement and final implant position (measured distances and angle with femur). The "Score" varied in the possible maximum score for each procedure and was therefore converted to a percentage of maximum score (PM score).

To explore the reliability and validity of the test, we included 20 physicians in the study (10 novices who were orthopedic interns with no prior surgical experience in operating hip fractures (group 1), and 10 orthopedic surgeons (senior residents or specialists) with experience of more than 20 hip fractures (group 2)). Recruitment was done in hospitals in the Capital Region of Denmark and all participation was voluntary. None of the participants had been trained on the simulator before the test. Physicians were tested between April 2013 and May 2013. Testing was done at Copenhagen University Hospital, Hvidovre and at Copenhagen University Hospital, Rigshospitalet. All the physicians were tested on the same simulator and in the same setting. The principal researcher (PP) supervised all the tests. During the "warm-up", PP was standing by to help with the simulator throughout the procedure. The test was completed immediately after the "warm-up" and PP administered it. PP operated the keyboard to help select requested screw length and drill length, as a nurse would help in the operation room. Necessary information such as "the available cortical screw lengths are 32–55 mm" was given during the procedures. To prevent bias, the information was written beforehand and was given in the same way to each participant.

Statistics

Independent-samples t-tests was used to compare the performance of the novice group and the experienced group. Each of the 5 simulator metrics described above was tested for statistically significant differences between the 2 groups. Levene's test for equality of variances was performed, and if equal variances could be assumed ($p > 0.05$), we used Student's t-test whereas Welch's t-test was used when equal variances could not be assumed ($p < 0.05$). Only valid simulator metrics (i.e. simulator metrics that could discriminate between the performances of the 2 groups) were analyzed further regarding inter-case reliability by calculating an intra-class correlation coefficient (ICC). The PM score for each procedure was combined to a mean PM score for the novices (group 1) and one for the experienced surgeons (group 2). The mean PM score distribution of the 2 groups was plotted using the contrasting-groups method (Downing and Yudkowsky 2009). The intersection between the distributions of the 2 groups was set as the pass/fail standard, and the consequences of the pass/fail standard were explored.

The statistical analysis was performed using SPSS version 19. Differences in metrics were considered to be statistically significant when the p-value was < 0.05.

Results

The combined score, expressed as the PM score, showed statistically significant differences between the novices and the experienced surgeons, whereas none of the individual simulator metrics demonstrated discriminatory abilities (Table). The inter-case reliability of the PM score was 0.83 (95% CI: 0.65–0.93) (ICC, average measures). The ICC for single measures was 0.62 (95% CI: 0.38–0.81). The mean PM score for the novices was 30% (SD 32, CI: 7.3–53) and for the experienced surgeons it was 76% (SD 10, CI: 68–83) ($p < 0.001$). The pass/fail standard was classidied as a mean PM score of 58% (Figure 1). All the novices and a single experienced surgeon failed the test (Figure 2).

Discussion

We found that a combination of simulator metrics—in this study expressed as the PM score for each procedure—was the only measure to show discriminative ability, whereas the 4 individual simulator metrics were similar between novices and experienced surgeons.

Demographics and procedure performances for novices and experienced surgeons on a virtual-reality hip-surgery simulator. Values are mean (SD)

	Novice (n = 10)	Experienced (n = 10)	p-value
Median age (range)	29 (26–39)	36 (30–58)	
Sex (F/M)	5/5	0/10	
Procedure 1 (Cannulated screws)			
Fluoroscopy time in seconds	55 (50)	21 (10)	0.06
No. of radiographs	73 (61)	76 (22)	0.9
No. of retries in guide placement	0.4 (0.5)	1.3 (1.3)	0.07
Procedure time in seconds	429 (88)	434 (91)	0.9
Percentage of maximum (PM) score	34 (32)	74.3 (16)	0.002
Procedure 2 (Hansson pin)			
Fluoroscopy time in seconds	44 (50)	19 (8)	0.1
No. of radiographs	79 (59)	75 (38)	0.9
No. of retries in guide placement	0.8 (1.0)	1.9 (2.3)	0.2
Procedure time in seconds	382 (116)	393 (128)	0.8
Percentage of maximum (PM) score	39 (35)	78 (9.8)	0.003
Procedure 3 (dynamic hip screw)			
Fluoroscopy time in seconds	70 (100)	18 (14)	0.1
No. of radiographs	85 (48)	64 (23)	0.2
No. of retries in guide placement	1.1 (1.5)	1.5 (1.3)	0.5
Procedure time in seconds	501 (172)	377 (81)	0.05
Percentage of maximum (PM) score	18.4 (53)	75.3 (13)	0.008

However, previous studies have indicated that some of the individual simulator metrics may have discriminatory abilities. Tillander et al. (2004) explored simulator metrics for distal femoral nailing and found statistically significant differences regarding procedure time and numbers of radiographs. This may be explained by the level of expertise in the groups tested. In our study, the novices group was orthopedic interns whereas Tillander et al. (2004) used medical students. In a study by Froelich et al. (2011), discriminative ability was found regarding fluoroscopy time and retries regarding the dynamic hip screw application. 15 residents were divided into 2 groups based on their years of postgraduate training. We could not reproduce these results, which may mean that we had type-II error. Even though their groups were smaller, they tested their residents 6 times repeatedly in the same procedure whereas we tested once in 3 different procedures. Assessment of more procedures in a test will increase the reliability but will reduce the feasibility in terms of test duration (Konge et al. 2011).

Interestingly, experienced surgeons had a tendency to use more attempts to place the guide. The most likely explanation is that experienced surgeons may be more determined to obtain the optimal placement of the K-wire guide due to awareness of the fact that a suboptimal implant position is a known predictor of later fixation failure (Baumgaertner and Solberg 1997). Using more attempts in placing the guide may result in use of more time, X-rays and fluoroscopy in the first part of each procedure. Unfortunately, the simulator metrics did not allow us to distinguish between the different sections of the procedures. It is possible that differences between novices and experienced surgeons were evened out through the sections of each procedure that followed. The results obtained in a simulation setting can probably not be directly translated to the clinical setting, in which a supervising surgeon would be present and interfere in case of an incorrect placement of the guide. Novices might therefore use more fluoroscopy when being supervised in operating on patients (Giannoudis et al. 1998).

A test reliability of > 0.8 is necessary for important decisions such as deciding when a trainee is ready to perform supervised operations on patients (summative assessment) (Downing and Yudkowsky 2009). Our study showed that assessment of a single procedure did not meet this criterion

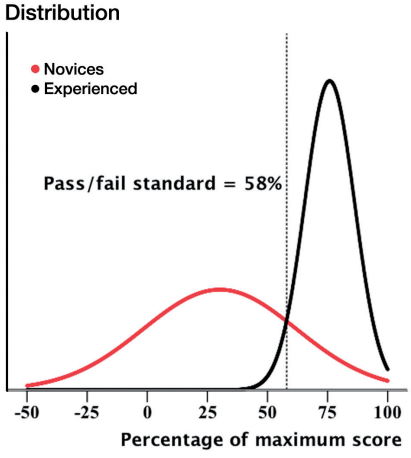

Figure 1. Distribution of percentage of maximum (PM) score for novices and experienced surgeons. Using the contrasting-groups method, the pass/fail standard for the test was determined from the intersection of the distributions (58%).

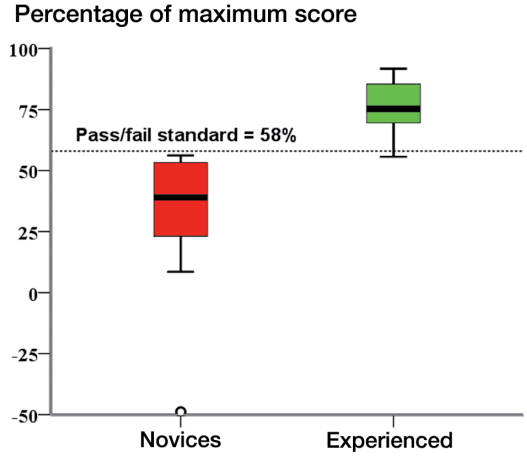

Figure 2. Box-plot showing percentage of maximum score (PM score) for novices and experienced surgeons, respectively. The line illustrates the consequence of the pass/fail standard. None of the novices passed the test. One of the experienced surgeons also failed the test.

(ICC 0.62), but combining 3 procedures into 1 test ensured the necessary reliability (ICC 0.83). This underpins the importance of testing in multiple procedures. The PM score showed discriminative ability between novices and experienced surgeons, which counts as evidence of construct validity (Ringsted et al. 2011). The standard deviation of PM scores for novices was high, mainly due to a single outlier with a negative score as seen in Figure 2. A negative score can occur when the operator severely damages the patient, e.g. by fracturing more bone. Figure 2 indicates that the experienced surgeons performed more consistently than the novices, which corresponds well with the Fitts and Posner model of development of motor skills—where consistency is a trademark of experts in the final, autonomous, stage of learning motor skills (Magill 2007).

There is no gold standard for the standard-setting regarding pass/fail score, but credible standards have to be established based on respected methods (Konge et al. 2013). We used the contrasting-groups method, and our greatest concern was whether the novices with no prior surgical experience in operating hip fractures would be able to pass. With a pass/fail standard of 58% in PM scores, none of the novices passed but 1 experienced surgeon failed. With the relatively wide distribution of the novices' scores, it could be necessary to increase the pass/fail standard cutoff for passing even more (Magill 2007). Stefanidis et al. (2012b) argued that trainees who achieved expert levels showed more automaticity and were safer in the operating room. Increasing the pass/fail standard beyond 75% could ensure the skills of simulator-trained novices, but would also lead to the failing of more experienced surgeons.

In Denmark, internal fixation of hip fractures is one of the first operations that orthopedic residents perform. 5 procedures must be completed in the first year of training (Frederiksen 2010). For simulation-based training to have an impact on future surgeons' education, it has to be integrated in the national curriculum and the learning outcome would have to be assessed (Scott and Dunnington 2007, Downing and Yudkowsky 2009). In future studies, it would be interesting to explore the effects of incorporating a simulation-based training program for residents (Karam et al. 2013). Such a training program could use the test we present as an end-of-course examination (training to criterion). Necessary and appropriate transfer studies should be undertaken to measure the effect regarding improvement of clinical surgical skills and improved patient outcome.

The present study had several limitations. The sample sizes were small, but comparable to similar studies on virtual-reality simulators (Tillander et al. 2004, Froelich et al. 2011, Konge et al. 2013). Furthermore, it is important to acknowledge the fact that the simulator only allows us to assess technical skills. Other important aspects, such as the ability to set the correct indication for surgery, should still be tested using direct observation in the real world.

In summary, we found it feasible to combine the 3 procedures for internal fixation of hip fractures on the STV simulator into a reliable and valid test. Performance of 3 procedures ensured a reliability of > 0.8, and a credible pass/fail standard could be determined. The test and the pass/fail standard could help assess and guarantee the quality of future trainees in simulation-based training programs before they proceed to supervised practice on patients.

Planning and design of the study: PP, HP, CR, and LK. Hypothesis: PP and LK. Statistical analysis: LK. Writing of the manuscript: PP. All the authors revised and approved the final manuscript.

There was no external funding for the study. The study was performed completely independently of the simulator company. None of the authors have any competing interests to declare.

Atesok K, Mabrey J D, Jazrawi L M, Egol K A. Surgical simulation in orthopaedic skills training. J Am Acad Orthop Surg 2012; 20 (7): 410–22.

Baumgaertner M R, Solberg B D. Awareness of tip-apex distance reduces failure of fixation of trochanteric fractures of the hip. J Bone Joint Surg (Br) 1997; 79 (6): 969–71.

Blyth P, Stott N S, Anderson I A. A simulation-based training system for hip fracture fixation for use within the hospital environment. Injury 2007; 38 (10): 1197–203.

Blyth P, Stott N S, Anderson I A. Virtual reality assessment of technical skill using the Bonedoc DHS simulator. Injury 2008; 39 (10): 1127–33.

Cook D A, Hatala R, Brydges R, Zendejas B, Szostek J H, Wang A T, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA 2011; 306 (9): 978–88.

Downing S M, Yudkowsky R. Assessment in health professions education. Routledge 2009: 57-73.

Frederiksen B. Portefoelje_for_Introduktionsuddannelsen_rev__okt_10. 2010 Sep.

Froelich J M, Milbrandt J C, Novicoff W M, Saleh K J, Allan D G. Surgical simulators and hip fractures: A role in residency training? JSURG. Elsevier Inc; 2011; 68 (4): 298–302.

Giannoudis P V, McGuigan J, Shaw D L. Ionising radiation during internal fixation of extracapsular neck of femur fractures. Injury 1998; 29 (6): 469–72.

Karam M D, Pedowitz R A, Natividad H, Murray J, Marsh J L. Current and future use of surgical skills training laboratories in orthopaedic resident education: a national survey. J Bone Joint Surg (Am) 2013; 95 (1): e4.

Konge L, Arendrup H, Buchwald von C, Ringsted C. Using performance in multiple simulated scenarios to assess bronchoscopy skills. Respiration 2011; 81 (6): 483–90.

Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C. Using virtual-reality simulation to assess performance in endobronchial ultrasound. Respiration 2013; 86 (1): 59-65.

Leblanc J, Hutchison C, Hu Y, Donnon T. A comparison of orthopaedic resident performance on surgical fixation of an ulnar fracture using virtual reality and synthetic models. J Bone Joint Surg (Am) 2013; 95 (9): e601–6.

Mabrey J D, Reinig K D, Cannon W D. Virtual reality in orthopaedics: Is it a reality? Clin Orthop 2010; (468) (10): 2586–91.

Magill R A. Motor learning and control. 8 ed. New York, McGraw-Hill 2007 263-89.

Palm H, Jacobsen S, Krasheninnikoff M, Foss N B, Kehlet H, Gebuhr P. Influence of surgeon's experience and supervision on re-operation rate after hip fracture surgery. Injury 2007; 38 (7): 775–9.

Rambani R, Viant W, Ward J, Mohsen A. Computer-assisted orthopedic training system for fracture fixation. J Surg Educ 2013; 70 (3): 304–8.

Ringsted C, Hodges B, Scherpbier A. 'The research compass': An introduction to research in medical education: AMEE Guide No. 56. Med Teach 2011; 33 (9): 695–709.

Scott D J, Dunnington G L. The New ACS/APDS skills curriculum: Moving the learning curve out of the operating room. J Gastrointest Surg 2007; 12 (2): 213–21.

Stefanidis D, Arora S, Parrack D M, Hamad G G, Capella J, Grantcharov T, et al. Research priorities in surgical simulation for the 21st century. AJS. Elsevier Inc; 2012a; 203 (1): 49–53.

Stefanidis D, Scerbo M W, Montero P N, Acker C E, Smith W D. Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training. Ann Surg 2012b; 255 (1): 30–7.

Tillander B, Ledin T, Nordqvist P, Skarman E, Wahlström O. A virtual reality trauma simulator. Med Teach 2004; 26 (2): 189–91.