ORIGINAL REPORTS

Vessel Ligation Fundamentals: A Comparison of Technical Evaluations by Crowdsourced Nonclinical Personnel and Surgical Faculty

Yinin Hu, MD, Boxiang Jiang, MD, Helen Kim, BA, Anneke T. Schroen, MD, MPH, Philip W. Smith, MD and Sara K. Rasmussen, MD, PhD

Department of Surgery, University of Virginia School of Medicine, Charlottesville, Virginia

BACKGROUND: Evaluation of fundamental surgical skills is invaluable to the training of medical students and junior residents. This study assessed the effectiveness of crowdsourcing nonmedical personnel to evaluate technical proficiency at simulated vessel ligation.

STUDY DESIGN: Fifteen videos were captured of participants performing vessel ligation using a low-fidelity model (5 attending surgeons and 5 medical students before and after training). These videos were evaluated by nonmedical personnel recruited through Amazon Mechanical Turk, as well as by 3 experienced surgical faculty. Evaluation criteria were based on Objective Structured Assessment of Technical Skills (scale: 5-25). Results were compared using Wilcoxon signed rank-sum and Cronbach's alpha (α).

RESULTS: Thirty-two crowd workers evaluated all 15 videos. Crowd workers scored attending surgeon videos significantly higher than pretraining medical student videos (20.5 vs 14.9, $p < 0.001$), demonstrating construct validity. Across all videos, crowd evaluations were more lenient than expert evaluations (19.1 vs 14.5, $p < 0.001$). However, average volunteer evaluations correlated more strongly with average expert evaluations ($\alpha = 0.95$) than the strength of correlation between any 2 individual expert evaluators ($\alpha = 0.72$-0.88). Combined reimbursement for all workers was $80.00.

CONCLUSION: After adjustments for score inflation, crowdsourced can evaluate surgical fundamentals with excellent validity. This resource is considerably less costly and potentially more reliable than individual expert evaluations. (J Surg Ed ▮:▮▮▮-▮▮▮. © 2017 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: crowdsourcing, surgical simulation, surgical education, vessel ligation

COMPETENCIES: Patient Care, Practice-Based Learning and Improvement

INTRODUCTION

One of the principal aims of surgical education in the university hospital setting is the provision of safe and effective technical training to surgery novices. Owing to work hour restrictions and increased scrutiny of surgical outcomes, operative experiences at the bedside for students and residents have declined over the past decade.[1] In this environment, the application of simulation models to equip novices with technical fundamentals prior to entering the operating room has become commonplace.[2,3] For medical students, simulation curricula can increase the value of operating room learning experiences and nurture interests in surgery.[4,5]

When considering simulation curricula, the most common barriers are material and personnel costs. Although simulation systems vary drastically in cost and fidelity, there is evidence to support that—at least for fundamental surgical skills—low-fidelity bench-top simulators may be as efficacious as expensive biological and virtual-reality models.[6-9] In prior works, we established the construct validity and efficacy of a low-cost, low-fidelity vessel ligation model.[10,11] However, for this and many other low-cost benchtop simulation systems, the constant need for expert

Meeting presentation: American College of Surgeons Clinical Congress, Washington DC, October 2016 (Oral presentation).

Funding: This work was not supported by a third party outside of the institution. Institutional support was provided through the University of Virginia School of Medicine's Academy of Distinguished Educators.

PRECIS: Novices performing simulated vessel ligations were evaluated by faculty and crowdsourcing. Crowdsourced evaluations were found to have high reliability and validity for this fundamental surgical skill.

Correspondence: Inquiries to Sara K. Rasmussen, MD, PhD, Department of Surgery, University of Virginia School of Medicine, PO Box 800709, Charlottesville, VA 22908–0679; e-mail: skr3f@virginia.edu

evaluation remains a major obstacle to high-throughput, cost-effective training.

Two tools frequently implemented to assist technical training are machine learning algorithms and trained assistant instructors. Machine learning algorithms recognize patterns of proficient movements by calibrating against clusters of skilled performers.[12,13] When used appropriately, this approach can generate powerful virtual training tools such as those used in robotic surgery.[14,15] However, machine learning data inputs must be accurate and consistent in order to separate meaningful patterns from noise. For this reason, these algorithms may be suboptimal for evaluating hand movements in open surgery. Further, the calibration processes for these algorithms often requires large expert datasets, and may themselves be prohibitively costly. Owing to the inaccuracy of self-assessments among novices,[16-18] a number of studies have explored the use of trained assistant instructors in lieu of expert faculty to evaluate surgical fundamentals.[19,20] Effectiveness of this approach varies based on the level of experience of the assistant instructors. Although these instructors may be substantially less costly than faculty evaluators,[21] training clinically inexperienced instructors to evaluate a broad range of surgical skills in a consistent manner remains a challenge.

A more recent, inexpensive evaluative approach that has been facilitated by access to the Internet is crowdsourced assessment of technical skills (CSATS). This method uses large groups of clinically inexperienced survey responders to observe and evaluate fundamental skills. Advantages include quick response time, low cost, and broad applicability. However, there remain concerns regarding the validity of assessments generated by lay personnel with little or no medical background. In particular, when crowd responders assess only one or a few participants' techniques, there is little context through which to judge skill level. The purpose of this study is to assess the utility of CSATS in evaluating surgical fundamentals when workers are carefully selected and prompted to assess multiple performances of a single task. We hypothesized that, with this approach, CSATS would provide evaluations with high construct validity that correlate closely to those of expert evaluations.

METHODS

Data Collection

We recruited 6 fourth-year medical students and 6 experienced surgical faculty from the University of Virginia Department of Surgery to perform vessel ligations using a low-cost benchtop simulator. This simulator has previously been shown to be effective in both assessment and training settings.[11,22] In brief, the task simulates a laparotomy setting, and involves acquiring proximal and distal control on a synthetic vessel, followed by division and tie ligation using 2-0 silk ties. Upon enrollment, each participant observed 1 demonstration of the vessel ligation simulator. Participants then performed the vessel ligation task while being video-recorded such that only the distal wrist and hands were captured. Audio was not included in the video in order to minimize assessment bias based on perceived age and sex. For medical student participants, after an initial recorded attempt ("pretest"), a 45-minute one-on-one practice session was held involving both didactic instruction and at least 10 supervised practice attempts. These sessions were proctored by a mid-level general surgery resident (Y.H.). Within 1 week of completing this practice session, each student underwent a second recorded attempt ("posttest"). In this way, a total of 17 recordings (6 student pretests, 6 student posttests, and 5 faculty) were generated.

Evaluation

The online crowdsourcing market Amazon Mechanical Turk (Amazon.com Inc., Seattle, WA) was used to screen crowdsource workers for participation as evaluators. The screening process consisted of 3 initial videos of simulated vessel ligation: 1 experienced faculty, 1 fourth-year medical student pretest, and 1 negative control. The negative control comprised a vessel ligation video that was altered to introduce a prohibitive level of static. Workers were required to evaluate all 3 videos for acceptability of video quality, and to provide 5-point Likert scale ratings based on the Global Rating Scale (GRS) of the Objective Structured Assessments of Technical Skills (OSATS). Owing to the absence of an audio component and the workers' absence of medical background, only the following 5 OSATS subscores were measured: Respect for Tissue, Time and Motion, Instrument Handling, Flow of Operation, and Overall Performance. In order to pass the screening process, workers must indicate inadequate video quality for the negative control, and must rate the expert video a higher total score than the student video. For participation in screening, workers were each compensated $0.20.

Workers certified through the screening process were then recruited to evaluate a series of 15 videos in random order: 5 student pretests, 5 student posttests, and 5 faculty videos. The student and faculty videos used for screening were excluded from this dataset. Workers evaluated all 15 videos using the same OSATS global rating scale. For this more extensive evaluation process, each worker was compensated $1.00.

Three experienced surgical faculty were recruited to evaluate the same 15 videos as those provided to crowd workers. These expert evaluators reviewed the videos as their schedule availability allowed, and used the same grading scheme as that used by the crowd workers.

Analysis

Owing to nonnormal distribution, summary results are reported using median and interquartile range (IQR).