Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Crowd-sourced assessment of surgical skills in cricothyrotomy procedure

Nava Aghdasi, PhD,[a,*] *Randall Bly, MD,*[b] *Lee W. White, PhD,*[c] *Blake Hannaford, PhD,*[a] *Kris Moe, MD, FACS,*[b] *and Thomas S. Lendvay, MD*[d]

[a] *Department of Electrical Engineering, University of Washington, Seattle, Washington*
[b] *Department of Otolaryngology – Head and Neck Surgery, University of Washington, Seattle, Washington*
[c] *M.D.C. School of Medicine, Stanford University, Stanford, California*
[d] *Department of Urology, Seattle Children's Hospital Seattle, Washington*

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form
7 March 2015
Accepted 11 March 2015
Available online 18 March 2015

Keywords:
Surgical skill assessment
Emergency cricothyrotomy procedure
Crowd sourcing

ABSTRACT

Background: Objective assessment of surgical skills is resource intensive and requires valuable time of expert surgeons. The goal of this study was to assess the ability of a large group of laypersons using a crowd-sourcing tool to grade a surgical procedure (cricothyrotomy) performed on a simulator. The grading included an assessment of the entire procedure by completing an objective assessment of technical skills survey.

Materials and methods: Two groups of graders were recruited as follows: (1) Amazon Mechanical Turk users and (2) three expert surgeons from University of Washington Department of Otolaryngology. Graders were presented with a video of participants performing the procedure on the simulator and were asked to grade the video using the objective assessment of technical skills questions. Mechanical Turk users were paid $0.50 for each completed survey. It took 10 h to obtain all responses from 30 Mechanical Turk users for 26 training participants (26 videos/tasks), whereas it took 60 d for three expert surgeons to complete the same 26 tasks.

Results: The assessment of surgical performance by a group ($n = 30$) of laypersons matched the assessment by a group ($n = 3$) of expert surgeons with a good level of agreement determined by Cronbach alpha coefficient = 0.83.

Conclusions: We found crowd sourcing was an efficient, accurate, and inexpensive method for skills assessment with a good level of agreement to experts' grading.

1. Introduction

One challenge in surgical education is the accurate assessment of the technical skills of training surgeons. The current methods of surgical skills evaluation such as direct observation, assessment using Objective Assessment tools, or machine learning approaches mainly depend on the presence of senior surgeons [1]. They must observe the operation directly or watch the recorded videos offline to assign a grade, often by completing a survey such as objective structured assessment of technical skills (OSATS) [2]. Furthermore, the machine learning-based algorithms for skill assessments rely on labeled data for training, which must be accurate to generate meaningful models. Validating and labeling the data are not

* *Corresponding author*. Department of Electrical Engineering, University of Washington, 185 Stevens Way, Paul Allen Center - Room AE100R, Campus Box 352500, Seattle, WA. Tel.: +1 425 9857114; fax: +1 206-543-3842.
E-mail address: navaa@u.washington.edu (N. Aghdasi).
0022-4804/$ – see front matter
http://dx.doi.org/10.1016/j.jss.2015.03.018

Table 1 – Subjective postprocedure questionnaire.

Questions		Not at all	Somewhat	Neutral	Moderate	Very
		1	2	3	4	5
1	How realistic was the simulator?					
2	Was the model anatomically accurate?					
3	Did the model improve your understanding of how to perform the procedure?					
4	Did you find this simulator useful to practice the procedure?					

feasible without the collaboration of expert surgeons. Therefore, the objective assessment of the trainee surgeons may introduce a long interruption between performance and score for the students.

Crowd-sourced assessment of technical skill (CSATS) for assessing surgical performance is a novel, inexpensive method that can provide assessments in a short period of time. In this method, a crowd of individuals on the Internet who do not have specific training in surgical skills evaluate the surgeon's performance. Previously, crowd sourcing has been used for assessment of surgeons performing tasks such as robotic surgical suturing [3,4].

The goal of this study was to investigate the ability of the crowd to assess the performance of surgeons on a more complex procedure. A cricothyrotomy simulator was chosen as the test platform [5]. This procedure consists of multiple surgical subtasks such as making an incision with scalpel, spreading the tissue with forceps, retracting tissue with a cricoid hook, and inserting an endotracheal tube. In this article, we compare the performance scores given by expert surgeons to the scores given by crowd individuals.

2. Methodology

2.1. *Data collection*

After Institutional Review Board approval (University of Washington, IRB Number = 18,192), 26 participants comprised of medical students, residents, and attending surgeons from the University of Washington Department of Otolaryngology-Head and Neck Surgery, performed a cricothyrotomy procedure using a low-cost simulator [5]. The simulator was designed and developed in the BioRobotics Lab, University of Washington. The simulator emphasizes the palpation and the correct identification of anterior cervical anatomy. It is optimized for materials that are low-cost, widely available, and simple to assemble. To validate the simulator, the subjects were asked to complete the questionnaire about the simulator using 5-point Likert scale (Table 1). A subset of participants that were expert attending surgeons ($n = 12$) reported "moderate" or "very" related to accuracy (10/12) and realism (8/12) of the simulator and how it could be used as a training tool for improving the procedure knowledge (7/12) and practicing (10/12) this surgical procedure.

The participants were shown a brief slideshow about the procedure that culminated with a video tutorial published by the New England Journal of Medicine [6], and then their performance was video recorded for skills assessments. To make sure that the videos would not give away any information about the participants' experience levels, the videos were edited to only show the hands movement on the simulator. In addition, participants were asked to wear gloves while performing the procedure.

2.2. *Amazon Mechanical Turk evaluators*

Crowd-sourcing markets, such as Amazon Mechanical Turk, are online markets in which individuals post tasks ("human intelligence tasks-[HITs]") such as surveys, and workers are paid a small payment for completing these tasks.

An HTML survey form was created, which consisted of three parts. The first two parts were screening questions to ensure appropriate participation. In the first part, the individual observed a video of two surgeons, one novice and one expert, operating side by side on the simulator.

The Mechanical Turk workers were instructed to select the video with better performance. The goal of this section was to test the ability of the workers to differentiate between two very different skill levels. Because the difference between novice and expert skill is visually striking in the chosen examples, the first test is more about user attention and diligence than skill assessment qualification. This test may also filter out an automated robot attempting to complete the task. The second part of the survey was a paragraph of text, also designed to assess the workers attention level. In the assigned reading test, the subjects were instructed to not answer a subsequent question. Data from any user who answered the subsequent question was eliminated for not paying attention to the assigned task.

Finally, the workers were presented with a video of a participant performing the Cricothyrotomy procedure and were asked to grade the video by completing the OSATS questions. OSATS consisted of the following six questions, graded on a Likert (1 to 5) scale:

1. Respect for tissue.
2. Time and motion.
3. Instrument handling.
4. Knowledge of instruments.
5. Flow of operation.
6. Knowledge of the procedure.

Each question had five points. To assure that graders understood the scoring criteria, textual anchors were included in the survey and free text response fields were added to collect the workers' comments.

By using the Mechanical Turk web interface, we created a HIT for each survey. A *Hypertext Preprocessor* script collected the survey results and generated a unique survey code that the workers copied into the Mechanical Turk Web site for getting paid. Thirty workers were requested to complete each survey, and they were paid $0.50 for each successfully finished survey.

2.3. *Expert surgeon evaluators*

The videos of the participants performing the procedure (the same data as the third part of the crowd-source survey) were sent to three expert head and neck surgeons, faculty members in the Otolaryngology-Head and Neck Surgery Department at the University of Washington. Their assigned task was to complete the same OSATS-based questionnaire.

For each participant, the average score from the Mechanical Turk users was compared to the average score obtained from expert grading.

3. Results

Twenty-six HITs were created on Amazon Mechanical Turk, and all the responses (780 responses, 30 responses for each HITs) were obtained after 10 h. For three expert surgeons, it took 60 days to complete the video evaluations. Figure 1 shows the relationship between the surgeons' scores and crowd-sourced scores for each participant who performed the procedure, and 85% of the crowd-sourced scores fell within 5 points of the expert surgeons' scores. The paired *t*-test was calculated and showed that difference between the scores from two groups were not statistically significant. The $r = 0.833$ correlation coefficient and $r^2 = 0.7$ coefficient of determination demonstrate good correlation.

The score response distribution of both groups is shown in Figure 2, the average score given by crowd source was 19.24 (standard deviation = 5.6) compared with 19.77 (standard deviation = 3.2) given by the expert surgeons.

Cronbach alpha is the measure of internal consistency. Items that are internally consistent can be seen as raters that agree about the "true" value of the score associated with the subjects that participated in the experiment. In that sense, Cronbach alpha is an agreement coefficient, or measure of association. The level of agreement between the two graders (surgeons and Mechanical Turk workers) was 0.83, indicating a good agreement between two groups. Conventionally, alpha of >0.9 is considered excellent, >0.8 good, >0.7 acceptable, >0.6 questionable, >0.5 poor, and <0.5 unacceptable [7].

The score given to each participant by two groups is shown in Figure 3. The red bars indicate the cases where the difference between crowd source and surgeons' score is greater than 5 points. Moreover, Figure 3 indicates the subject classification into three groups (skilled, average, and beginner) based on their OSATS score (OSATS >20, 10< OSATS <20, and OSATS <10, respectively). The subjects were classified into the same skill level with both crowd source and surgeons except for two cases.

4. Discussion

Table 2 shows the 4 of 26 instances in which the crowd-sourced score deviates 5 points or more from the surgeons' score. In these occurrences, the variance of the surgeons' score is high too, suggesting that surgeons themselves were not in complete agreement. Only 2 of 4 participants with 5 scores apart were classified into different skill levels. Figure 3 results show that laypersons can identify highly skilled participants correctly, yet they may have difficulty to distinguish between average or beginner skilled participants. One possible explanation is that there was some variability in grading if a participant completed the procedure correctly, yet did so slowly or with inefficient movements or poor tissue handling. The end result was a successfully completed procedure, but there were numerous opportunities for improved technique.

The average score for the two groups was very similar (within 0.6); however, the standard deviation of the Mechanical Turk users was higher (5.6 *versus* 3.2). Lack of experience and differing levels of understanding after watching the training video in the layperson graders may explain this greater variation, when compared with the group of expert graders.

Different metrics such as paired *t*-test and Cronbach alpha coefficient showed that the assigned scores by senior

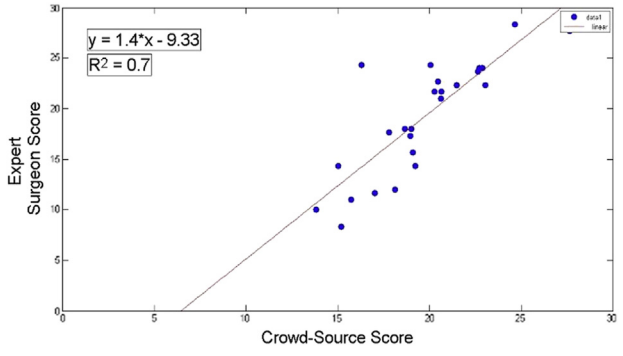

Fig. 1 – Surgeons score *versus* Mechanical Turk workers' score. (Color version of figure is available online.)

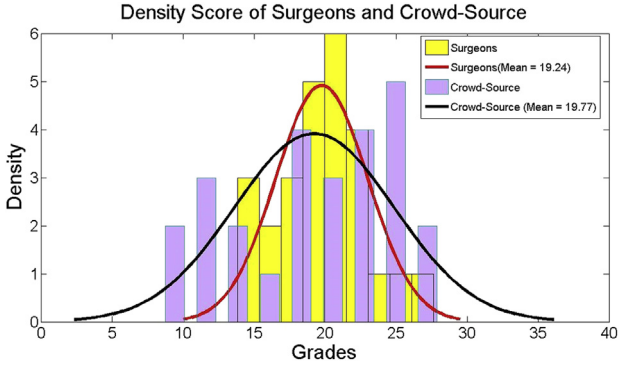

Fig. 2 – Score density for expert surgeons and Mechanical Turk workers. (Color version of figure is available online.)

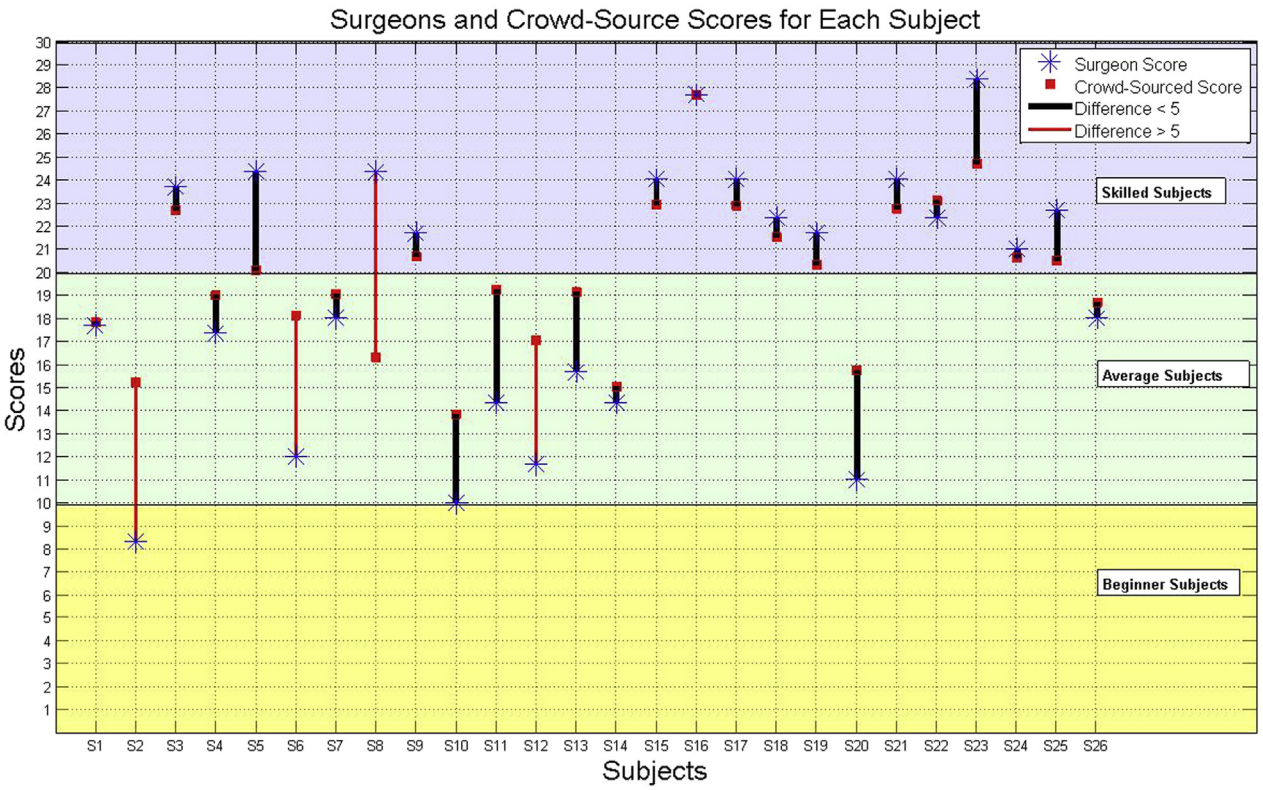

Fig. 3 – Score for each subject given by expert surgeons and Mechanical Turk workers. (Color version of figure is available online.)

surgeons correlate with those obtained from crowd sourced, but the scores were obtained in a significantly shorter period, and using less resources. Prior work has demonstrated accuracy in crowd-sourced assessment in simple tasks, but this study shows the same result with a more complex procedure with multiple subtasks [3,8].

Amazon allows the survey creator to set the price for each HIT. Crowd workers may put more time and effort into a task, which pays more. Thus, improved pay and more detailed instructions for workers could even further increase these agreements.

Increasing the pay for each HIT has other advantages too. For example, more questions could be added to the survey. Questions regarding demographic data of the crowd workers, which Amazon Mechanical Turk does not provide, can be used for further analysis of the responses.

The ability to provide rapid, scalable, objective performance assessment makes CSATS an interesting method because faculty time (for evaluating and grading) is a precious resource. We foresee that in resident and medical student simulation education, some of the self-directed surgical skills practice that is not directly observed may now be able to viewed and assessed for feedback. Although practice makes perfect, imperfect practice makes imperfect performance. CSATS may be used to triage trainees for targeted curricula as opposed to training all residents with the same limited resources.

As accreditation bodies such as the Accredited Council on Graduate Medical Education and the Residency Review Committee demand objective performance benchmarking and competency-based advancement, there will be an increasing need to solve the challenge of providing objective assessment of skill with finite faculty time. Using crowds to assist in skills evaluation may be part of the solution.

Furthermore, CSATS could have a utility in resource-poor learning environments. Future work will focus on determining the benefit of such a system in those areas. A major limitation to crowd-sourced feedback for surgical trainees is

Table 2 – Participants with more than 5-point difference in their assigned scores by two groups.

Participants	Surgeon 1 score	Surgeon 2 score	Surgeon 3 score	Variance surgeons' score	Crowd-sourced score
1	16	9	11	13	18
2	11	10	14	4.33	17
3	26	24	23	2.33	16.3
4	10	8	7	2.33	15.2

that it could compromise the essential role of surgical mentors. This method is not intended to replace timely feedback provided by the mentor at the time of the procedure but rather to augment current training methods. It will add an objective standardization and may have a future role in accreditation and competency assessment.

5. Conclusions

Crowd sourcing is an efficient, accurate, and inexpensive method for skills assessment, even when applied to a complex procedure. The assessment of surgical performance by a group ($n = 30$) of laypersons matched the assessment by a group ($n = 3$) of expert surgeons (correlation coefficient 0.833). These results have strong implications on surgical training, in both resource-rich and resource-poor environments and provides a unique quantitative assessment to a procedure that is otherwise graded qualitatively.

Acknowledgment

The authors thank Dr Isaac Bohannon and Dr Mark Whipple for their assistance in grading the cricothyrotomy videos. This research was sponsored by University of Washington Global Health Department–Seed Grant.

Authors' contributions: All authors contributed extensively to the work presented in this article. All authors contributed to the simulator design, experiment design, and data collections. All authors discussed and contributed to the survey design. All authors discussed the results and implications and commented on the article at all stages.

Disclosure

N.A., R.B., L.W.W., B.H., and K.M. report no proprietary or commercial interest in any product mentioned or concept discussed in this article. T.S.L. is equity owner of CSATS, Inc. This research preceded the commercialization of the CSATS method.

REFERENCES

[1] Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surg Endosc 2011;25:356.
[2] Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997;84:273.
[3] Chen C, White L, Kowalewski T, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 2014;187:65.
[4] Holst D, White L, Brand T, Harper J, Lendvay T. Crowd-sourced assessment of technical skills (C-SATS): an adjunct to urology resident simulation training. J Endourol; 2014 [Epub ahead of print].
[5] White L, Bly RA, D'Auria D, et al. Cricothyrotomy simulator with computational skill assessment for procedural skill training in the developing world. Otolaryngology–Head Neck Surg 2013;149(2 Suppl l):P60.
[6] Hsiao J, Pacheco-Fowler V. Cricothyroidotomy. N Engl J Med 2008;358:e25. Available from: http://www.nejm.org/doi/full/10.1056/NEJMvcm0706755.
[7] Gliem JA, Gliem RR. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Res Pract Conf Adult Contin Community Educ; 2003.
[8] White L. Quantitative objective assessment of preoperative warm-up for robotic surgery. Ph.D. dissertation. Seattle, WA: Dept. BioMed. Eng, University of Washington; 2013.