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.