Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Crowd-Sourced Assessment of Technical Skills: a novel method to evaluate surgical performance

Carolyn Chen, BA,[a] Lee White, PhD,[b] Timothy Kowalewski, PhD,[c] Rajesh Aggarwal, MD, PhD,[d] Chris Lintott, PhD,[e] Bryan Comstock, MS,[f] Katie Kuksenok, BA,[g] Cecilia Aragon, PhD,[g] Daniel Holst, BS,[a,] and Thomas Lendvay, MD[h]*

[a] *University of Washington, School of Medicine, Seattle, Washington*
[b] *Department of Bioengineering, University of Washington, Seattle, Washington*
[c] *Department of Mechanical Engineering, University of Minnesota, Seattle, Washington*
[d] *Department of Surgery, University of Pennsylvania, Philadelphia, Pennsylvania*
[e] *Department of Physics, University of Oxford, Oxford, United Kingdom*
[f] *Department of Biostatistics, University of Washington, Seattle, Washington*
[g] *Department of Computer Science Engineering, University of Washington, Seattle, Washington*
[h] *Department of Urology, University of Washington, Seattle Children's Hospital, Seattle, Washington*

ARTICLE INFO

Article history:
Received 31 July 2013
Received in revised form
6 September 2013
Accepted 18 September 2013
Available online 10 October 2013

Keywords:
Crowdsourcing
Robotic surgery
OSATS
GEARS
Education
Training

ABSTRACT

Background: Validated methods of objective assessments of surgical skills are resource intensive. We sought to test a web-based grading tool using crowdsourcing called Crowd-Sourced Assessment of Technical Skill.

Materials and methods: Institutional Review Board approval was granted to test the accuracy of Amazon.com's Mechanical Turk and Facebook crowdworkers compared with experienced surgical faculty grading a recorded dry-laboratory robotic surgical suturing performance using three performance domains from a validated assessment tool. Assessor free-text comments describing their rating rationale were used to explore a relationship between the language used by the crowd and grading accuracy.

Results: Of a total possible global performance score of 3–15, 10 experienced surgeons graded the suturing video at a mean score of 12.11 (95% confidence interval [CI], 11.11–13.11). Mechanical Turk and Facebook graders rated the video at mean scores of 12.21 (95% CI, 11.98–12.43) and 12.06 (95% CI, 11.57–12.55), respectively. It took 24 h to obtain responses from 501 Mechanical Turk subjects, whereas it took 24 d for 10 faculty surgeons to complete the 3-min survey. Facebook subjects (110) responded within 25 d. Language analysis indicated that crowdworkers who used negation words (i.e., "but," "although," and so forth) scored the performance more equivalently to experienced surgeons than crowdworkers who did not ($P < 0.00001$).

Conclusions: For a robotic suturing performance, we have shown that surgery-naive crowdworkers can rapidly assess skill equivalent to experienced faculty surgeons using

* *Corresponding author*. Department of Urology, University of Washington, Seattle Children's Hospital, 4800 Sand Point Way NE, Seattle, WA, PO Box 359300. Tel.: +1 307 752 1996; fax: +1 206 987 3925.
E-mail address: dholst12@gmail.com (D. Holst).

http://dx.doi.org/10.1016/j.jss.2013.09.024

Crowd-Sourced Assessment of Technical Skill. It remains to be seen whether crowds can discriminate different levels of skill and can accurately assess human surgery performances.

1. Introduction

The annual mortality because of medical errors may be as high as 98,000 patients in the United States [1]. Even more patients experience morbidity yielding consequences both clinically and economically [1]. An extra 2.4 million hospital days and $9.3 billion are incurred annually because of medical errors [2]. Efforts to reduce surgical complication rates have included incorporation of simulation training for learning and recertification of surgical skills [3]. Global surgical performance-rating scales, such as the Objective Structured Assessment of Technical Skills (OSATS), have been widely adopted for the assessment of surgical skill and the determination of trainee advancement [4,5]. These methods, although validated, are time-intensive and rely on real-time or video-recorded analysis by surgical experts who first need to demonstrate inter-rater reliability. Increasing responsibilities of surgical educators and the trend toward standardization of training dictate a need for a cheaper, faster, less biased method of rating surgical performance.

Crowdsourcing is a relatively recent trend that uses an anonymous crowd to complete small, well-defined tasks [6]. The crowd must be diverse, decentralized, and independent, and the generated data need to be able to be aggregated [7]. Ongoing research in the area investigates how to define tasks in a way that enable the crowd to accomplish complex and expert-level work. Various workflows [8] can be used to break a complex piece of work into approachable parts and can also use the crowd to check the quality of its own work [9]. Crowdsourcing has been used to help blind mobile phone users navigate their environment [10], decipher complex protein folding structures with the online game called *Foldit* [11], and solve medical cases through the website CrowdMed.com [12]. These applications all use online work marketplaces, such as Amazon Mechanical Turk [13] to quickly and cheaply recruit an anonymous crowd of nonexperts. We hypothesize that crowd-sourced surgery performance rating is equivalent to ratings done by experienced surgeons. We also explored a link between the language of the crowd and more accurate ratings of surgical performances.

2. Materials and methods

After Institutional Review Board approval (IRB #42,811), three groups of subjects were recruited for this study: Amazon.com Mechanical Turk users, Facebook users, and teaching surgeons whose expertise and practice involve robotic surgery. Recruitment emails to the experienced surgeons were sent and Mechanical Turk and Facebook announcements were posted on the respective websites. Five hundred one subjects were recruited through the Amazon.com Mechanical Turk crowdsourcing platform (https://www.mturk.com/mturk/welcome) (Fig. 1A). Eligible subjects were active Mechanical Turk users who had completed 50 or more Human Intelligence Tasks, the task unit used by Mechanical Turk, and had achieved a greater than 95% approval rating. Each Mechanical Turk subject was compensated $1.00 for participating. In the second group, 110 subjects were recruited using Facebook (Fig. 1B). The control group consisted of 10 experienced robotic surgeons, who have all practiced as attending surgeons for a minimum of 3 y with predominantly minimally invasive surgery practices and who were familiar with evaluating surgical performances by video analysis (Fig. 1C). Neither the Facebook subjects nor the surgeon raters received monetary compensation. All subjects were required to be older than 18 y.

A surgical skill assessment survey was adapted from the Global Evaluative Assessment of Robotic Skills (GEARS) validated robotic surgery rating tool [14] and hosted online (Fig. 2). Each of the subjects from the three groups completed the same survey. The survey consisted of two steps. First, the subjects were asked to answer a qualification question in which a pair of videos of surgeons performing a Fundamentals of Laparoscopic Surgery block transfer task were displayed side by side on the screen [15] (Fig. 3). These videos were obtained from a previous study [16]. The left video demonstrated a surgeon performing with high skill, whereas the right video presented a surgeon performing with intermediate skill based on published benchmark metrics for this particular task [17,18]. Subject assessors were directed to indicate which video showed the surgeon of higher skill. This question was used to assess the subject's discriminative ability. After the qualification question, the criterion test involved rating a less than 2-min robotic surgery suture knot-tying video of an above average performance (Fig. 4) based on existing benchmark data [17,18]. No subject-identifying features were visible. After watching the video, each reviewer rated the suturing performance on three domains: depth perception, bimanual dexterity, and efficiency (Fig. 2). The domains were chosen from the six domains included in the GEARS tool and were rated on a Likert scale from 1–5 [14]. The global performance rating was obtained by summing the ratings of the three domains with a scale of 3–15. An attention question was also embedded within the criterion test to ensure that the assessor was actively paying attention and if the question was answered incorrectly, the subject was excluded from the study.

The assessor was asked to describe his or her grading rationale in a free-text box after rating for each domain. We focused on using the occurrence of style words, which are words that do not carry content individually, such as "the," "and," "but," and "however," to identify more accurate responses. Chung and Pennebaker distinguished between content and style words in text analysis, and found that noncontent words in English can help identify aspects of the writer's mood, expertise, and other characteristics [19]. In an exploratory step, we split all qualifying responses into two groups: those closer to the expert answers, and those farther