17-7123

Is Crowdsourcing Surgical Skill Assessment Reliable? An Analysis of Robotic Prostatectomies

Thomas S. Lendvay, Seattle, WA, Khurshid R. Ghani, Ann Arbor, MI, James O. Peabody, Detroit, MI, Susan Linsell, David C. Miller, Ann Arbor, MI, Bryan Comstock, Seattle, WA*

INTRODUCTION AND OBJECTIVES: Crowdsourcing has demonstrated the ability to provide accurate surgical skills assessments correlating with expert surgeon reviewers. We studied whether crowdsourcing skills assessment of robotic prostatectomy performances would yield reliable scoring across a range of days and times of day. We also sought to characterize the agreement of video review among peer robotic surgeons reviewing the same urethrovesical anastomosis videos.

METHODS: We used five urethrovesical anastomosis videos previously assessed by faculty-level surgeons within the Michigan Urological Surgery Improvement Collaborative (MUSIC) using the Global Evaluative Assessment of Robotic Skills (GEARS), (highest score 25). The bottom and top scoring videos and three videos evenly distributed across a range of peer rater scores were selected from a larger pool of videos. Each video was assessed through the C-SATS platform (C-SATS, Inc. Seattle, WA) by n=32 random crowdworkers at one of ten different days/times over one week. A C-SATS GEARS average score was generated for each video; reliability was assessed using intraclass correlation coefficient (ICC). We then evaluated the 5 anastomosis videos with n=23 faculty experts as a comparative example of expert rating performance. Experts were not subjected to the same repeated trials of reviews. Expert reviewers saw each performance once and responded within 14 days of the review process. Ten different groups of crowds reviewed each video over the course of the week.

RESULTS: A total of 342 unique crowdworkers provided 1,640 ratings in a median completion time of 1 hour and 22 minutes for each of the ten review sessions. The C-SATS ICC was found to be 0.92 (95% CI: 0.79 to 0.99), indicating a very high level of reliability of the C-SATS rating process (Figure 1a). Reliability was lower for the 23 faculty experts (ICC=0.68; 95% CI: 0.42 to 0.95), with the range of expert scores spanning more than half of the GEARS scale on each video (Figure 1b).

CONCLUSIONS: We demonstrated that crowdsourcing to assess technical skills is repeatable and reliable across multiple times of the week providing evidence that such a method could be used to assess the skill of surgeons. Furthermore, expert video review could potentially be enhanced through repeated trials with workshops to build consensus on scoring standardization.

Source of Funding: Blue Cross Blue Shield of Michigan

Figure 1a: Average C-SATS GEARS ratings on 5 prostatectomy anastomosis videos from n=32 pre-qualified reviewers taken at 10 different day-times. Figure 1b: GEARS scores from the same 23 faculty experts on the same 5 procedures. Each point in Figure 2b represents the GEARS score from an individual faculty expert.