Clinical Science

Can endoscopists accurately self-assess performance during simulated colonoscopic polypectomy? A prospective, cross-sectional study

James Ansell, B.Sc., M.B.B.Ch. a,*, Joanna J. Hurley, M.B.B.Ch. b, James Horwood, M.S. a, Chantelle Rizan, B.Sc. c, Konstantinos Arnaoutakis, M.Phil. a, Stuart Goddard, B.Sc. a, Neil Warren, Ph.D. a, Jared Torkington, M.S. d

aWelsh Institute for Minimal Access Therapy, Cardiff CF14 4UJ, UK; bUniversity Hospital Llandough, Cardiff, UK; cCardiff University, School of Medicine, Cardiff, UK; dUniversity Hospital of Wales, Cardiff, UK

KEYWORDS: Simulation; Endoscopy; Polypectomy; Training; Self-assessment

Abstract

BACKGROUND: The aim of this study was to establish if endoscopists can reliably self-assess their ability to perform simulated colonic polypectomy.

METHODS: Novices, intermediates, advanced, and experts performed a video-recorded polypectomy task using the Welsh Institute for Minimal Access Therapy (WIMAT) colonoscopy suitcase simulator. This involved removal of a simple polyp (A) and a complex polyp (B). Participants self-assessed themselves using a Direct Observation of Polypectomy Skills (DOPyS) assessment form. Two blinded, independent, Joint Advisory Group on Gastrointestinal Endoscopy (JAG) accredited assessors graded each performance using the same DOPyS scoring. The Spearman coefficient was used to determine the correlation between self and assessors’ scores.

RESULTS: Eighty participants completed the task. There was a weak correlation between assessors’ scores and self-assessment scores for all groups (novices: $r = .44, P = .85$; intermediates: $r = .16, P = .51$; advanced: $r = .16, P = .50$; and experts: $r = .07, P = .76$). There was a strong correlation between scores from assessor 1 and 2 for polyp A ($r = .80, P < .01$) and polyp B ($r = .80, P < .01$).

CONCLUSIONS: The correlation between self-assessment and assessors’ scores is weak. Novices and intermediates underestimate performance, whereas advanced and experts overestimate performance. Regular feedback may improve accuracy.

© 2014 Elsevier Inc. All rights reserved.

The ability to accurately self-assess performance is an important component of medical education. This is particularly true in practical specialties such as gastroenterology and surgery. In the United Kingdom, the Joint Advisory Group on Gastrointestinal Endoscopy (JAG) recommends that trainees use a personal development plan in order to highlight learning needs. Part of this process requires a degree of self-assessment and trainer-based formative review. This can be used to identify strengths for development and highlight weaknesses for correction.

Despite the proposed benefit, several studies from a variety of disciplines consistently indicate that the accuracy of self-assessment is poor. The majority of research focuses on written and clinical knowledge rather than technical skills. The evidence is contradictory when practical
tasks are considered. Moorthy et al found that senior surgical trainees are accurate in their self-assessment of technical procedures in a simulated operating room, whereas Pandey et al showed a poor correlation between self-appraisal and expert appraisal. Other articles report that high self-belief does not predict success, and in novices it corresponds negatively with skill.

The benefit of self-assessment has yet to be determined for therapeutic colonoscopy. Technically difficult tasks such as polypectomy, endoscopic mucosal resection, and endoscopic submucosal dissection are increasing in response to early polyp detection from bowel screening programs. The aim of this study was to determine if endoscopists of differing levels of expertise can accurately self-assess their performance in polypectomy using a novel simulator, the Welsh Institute for Minimal Access Therapy (WIMAT) colonoscopy suitcase. This is an ex vivo, porcine simulator validated for colonic polypectomy skills (WIMAT) colonoscopy suitcase. The same endoscopic equipment for therapeutic colonoscopy. Technically difficult tasks such as polypectomy, endoscopic mucosal resection, and endoscopic submucosal dissection are increasing in response to early polyp detection from bowel screening programs.

Methods

Participants

Eighty participants were recruited from the Wales Deanery, UK. A questionnaire established previous expertise and allowed each candidate to be assigned to 1 of the following groups: novices: junior doctors with no experience of colonoscopy, intermediates: specialist trainees (specialty training levels 3 to 7) with experience of performing more than 200 colonoscopies, advanced: JAG-accredited, independent colonoscopists; and experts: JAG-accredited bowel screening colonoscopists. For JAG accreditation (full certification), a trainee must be competent across several performance indicators including demonstrating a cecal intubation rate of more than 90% and a serious complication rate of less than .5%. To become a bowel screening colonoscopist, an independent practitioner must pass an additional written and practical assessment.

Study procedure

The WIMAT colonoscopy suitcase was used to perform 2 snare polypectomy procedures. This is a previously validated task, and the exact positioning of the polyps and setup of the simulator is described in earlier work (Fig. 1). The task was designed to allow trainees to perform a simple polypectomy (polyp A) and a complex polypectomy (polyp B). The setup of each simulation was identical, and the same endoscopy assistant was present for all cases. The luminal view of all procedures was video recorded for future performance analysis.

Assessment

After the polypectomy task, each participant completed a modified self-assessed Direct Observation of Polypectomy Skills (DOPyS) JAG form. This is marked on a scale of 1 to 4 as follows: 1: accepted standards are not yet met with frequent uncorrected errors; 2: some standards are not yet met, aspects to be improved, and some uncorrected errors; 3: competent and safe throughout procedure and no uncorrected errors; and 4: highly skilled performance. Any incomplete parameters are awarded a score of 0. Eight of 34 DOPyS parameters were selected according to their relevance to generic skills and for stalked polypectomy. DOPyS parameters were excluded if they could not be assessed by video format.

Two JAG-accredited colonoscopists reviewed each video and scored the performance using the same DOPyS assessment. Both assessors were given DOPyS descriptor guides for generic and stalked polyps. Assessors remained blinded to the level of experience of the participant at all times.

Statistical analyses

Previous research has shown that experts (equivalent to our definition of experts) have an 88% (n = 15) chance of scoring 3 to 4 (or pass) on the DOPyS, whereas nonexperts (equivalent to our definition of intermediate level) have a 53% (n = 8) chance of scoring between 3 and 4 (or pass) on the DOPyS. Assuming that the novice group would take a similar drop in performance (from 53% to 18%), then 20 participants in each group would give a greater than 80% power to detect a difference in performance of 35% between the groups using a 2-tailed test with a confidence level of 5%.

All statistical analyses were performed using the PASW Statistics Package 18, Hong Kong, China. An assessor’s score for each of the 8 DOPyS parameters was calculated by averaging the score of polyp A and B across both assessors. Inter-rater reliability of assessors and the relationship between self-assessment and expert assessment were determined using Spearman ρ correlation coefficients where less than .30 was considered to be a weak correlation, between .30
and .50 was considered a moderate correlation, and greater than .50 was considered a strong correlation. To complement this analysis, a Wilcoxon signed rank test was used to assess differences in median group performance. Data are expressed as medians with an interquartile range (IQR); a $P$ value less than .05 was considered statistically significant.

**Results**

Eighty participants (20 per group) attempted the polypectomy task, and all completed the self-assessment process after the simulation. Seven (8.8%) novices failed to complete polypectomy A, and 6 (7.5%) novices and 1 (1.3%) intermediate failed to complete polypectomy B. A statistically significant difference was shown between the groups for median overall competency assessment scores (novices vs intermediates, $P < .01$; novices vs advanced, $P \leq .01$; novices vs experts, $P \leq .01$; intermediates vs advanced, $P \leq .01$; intermediates vs experts, $P \leq .01$; and advanced vs experts, $P \leq .01$).

When the assessors’ scores were compared with the self-assessment scores of the total cohort of 80 participants, the correlation was strong ($\rho = .70$, $P \leq .01$, Fig. 2). However, when the groups were considered separately, the correlation was weak. The novice median assessment score for overall competency was 1.00 (IQR = .50 to −1.25) compared with a self-assessment score of .50 (IQR = .00 to 1.00) ($\rho = −.44$, $P = .85$). For intermediates, the overall competency assessors’ score was 2.37 (1.75 to 2.50), and the self-assessment score was 2.00 (1.00 to 2.00) ($\rho = −.16$, $P = .51$). In the advanced group, the overall competency assessors’ score was 2.75 (2.50 to 3.25), and the self-assessment score was 3.00 (3.00 to 3.00) ($\rho = .16$, $P = .50$).

No statistically significant difference was observed between advanced vs experts, $P = .14$. Significant differences were also demonstrated in the self-assessment median group scores for overall competency (novices vs intermediates, $P \leq .01$; novices vs advanced, $P \leq .01$; intermediates vs advanced, $P \leq .01$; and intermediates vs experts, $P \leq .01$).

Three DOPyS parameters measured showed a moderate to strong correlation for overall competency in polypectomy (novices [polyp A: $\rho = .62$, $P \leq .01$ and polyp B: $\rho = .50$, $P = .03$], intermediates: [polyp A, $\rho = .30$, $P = .30$ and polyp B: $\rho = .60$, $P \leq .01$], advanced [polyp A: $\rho = .50$, $P \leq .01$; intermediates vs advanced, $P \leq .01$; intermediates vs experts, $P \leq .01$; and advanced vs experts, $P \leq .01$].

Inter-rater reliability between the assessors showed a moderate to strong correlation for overall competency in polypectomy (novices [polyp A: $\rho = .62$, $P \leq .01$ and polyp B: $\rho = .50$, $P = .03$], intermediates: [polyp A, $\rho = .30$, $P = .30$ and polyp B: $\rho = .60$, $P \leq .01$], advanced [polyp A: $\rho = .50$, $P \leq .01$; intermediates vs advanced, $P \leq .01$; intermediates vs experts, $P \leq .01$; and advanced vs experts, $P \leq .01$].

**Figure 1** The WIMAT colonoscopy suitcase. (A) The laparoscopic entry port for the colonoscope, (B) the outer casing of the simulator, (C) the internal view of the simulator with the bowel opened to display the inserted polyps, (D) the position of polyp A (simple) inserted at 25 cm from the anal verge in the 6 o’clock position in front of a luminal fold, (E) the position of polyp B (complex) inserted at 43 cm from the anal verge and positioned at 1 o’clock behind a luminal fold and distal to a simulated colonic bend. 1 = polyp A; 2 = polyp B.
When an overall group comparison was performed for all groups together, the correlation was strong for polyp A (\( r = .80, P < .01 \), Fig. 3) and polyp B (\( r = .80, P < .01 \), Fig. 4).

**Comments**

This is the first study to evaluate the reliability of self-assessment during simulated colonoscopic polypectomy. There is a lack of consistency in the literature to conclude whether accurate self-assessment of technical skill is possible. A meta-analysis of 44 self-assessment studies in higher education reported a moderate correlation between self- and expert assessments of .39.\(^{15}\) A similar review by Gordon\(^{3}\) of 18 articles showed comparable findings. However, when medical trainee self-assessments are compared with expert scores, the correlation is usually weak.\(^{8,16–18}\)

Our results for trainee self-assessment in simulated colonoscopic polypectomy are consistent with these findings. The relationship between self-assessment and independent assessment of advanced and expert colonoscopists again shows a weak statistical correlation. This is contrary to some reports in the literature that the ability to accurately self-assess improves with experience because the participant can recognize an expert performance and use this as a benchmark to assess his/her own skills.\(^{7,19}\) Despite a weak statistical relationship, self-assessment scores did in general increase with real-life levels of expertise. Novices awarded themselves lower self-assessment scores than the intermediates and the intermediates lower scores than the advanced and experts.

Novices and intermediates tended to underestimate their ability (median assessors’ score for novices = 1.00 [IQR .50 to 1.25] and 2.38 (IQR 1.75 to 2.50) for intermediates compared with a self-assessment median score for novices of .50 [IQR .00 to 1.00] and 2.00 [IQR 1.00 to 2.00] for intermediates). On the other hand, the advanced and experts groups overestimated their ability (median assessor score for advanced = 2.75 [IQR 2.50 to 3.25] and 3.00 (IQR 2.75 to 3.44) compared with self-assessment median scores of 3.0 [IQR 3.0 to 3.0] for advanced and 4.00 [IQR 3.00 to 4.00] for experts). These findings can be interpreted in several ways. First, trainees lack experience of performing real-life polypectomy and are therefore less able to accurately self-assess their performance. Advanced and expert colonoscopists may be familiar with assessing novice and intermediate performances but may be less able to repeat this process for themselves. They may also feel pressured to represent themselves in the best possible light. This term is known as impression management and has been previously applied to trainees but may, in fact, also be true of trainers themselves.\(^{20}\)

Alternatively, the over- or underestimation can be directed at the assessment process itself. There are currently no valid, quantitative measures of polypectomy assessment. The DOPyS score is a subjective measure of skill that may be open to interpretation. It has been validated for the assessment of a range of real-life polypectomy procedures but has not been used in simulated polypectomy assessment previously. This may affect the reliability of its use in this trial. However, 2 blinded,
## Table 1  Summary of assessors’ scores versus self-assessment scores for all groups and for each DOPyS parameter used

<table>
<thead>
<tr>
<th>Group</th>
<th>Score</th>
<th>DOPyS parameter assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assessor score, median (IQR)</td>
<td>Achieves optimal position</td>
</tr>
<tr>
<td>Novices</td>
<td>Self-assessment, median (IQR)</td>
<td>Optimizes view by aspiration/insufflation/wash</td>
</tr>
<tr>
<td>(n = 20)</td>
<td>Spearman ρ P value</td>
<td>Directs snare accurately over polyp head</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Places the snare at appropriate position on the stalk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ensures appropriate amount of tissue trapped</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Applies appropriate degree of diathermy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Retrieves or attempts for retrieval of polyp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall competency at polypectomy</td>
</tr>
<tr>
<td>Novices</td>
<td>Assessor score, median (IQR)</td>
<td>1.00 (.75–1.25)</td>
</tr>
<tr>
<td>(n = 20)</td>
<td>Self-assessment, median (IQR)</td>
<td>.50 (.00–1.00)</td>
</tr>
<tr>
<td></td>
<td>Spearman ρ P value</td>
<td>−.71 P = .77</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Assessor score, median (IQR)</td>
<td>2.00 (1.75–2.69)</td>
</tr>
<tr>
<td>(n = 20)</td>
<td>Self-assessment, median (IQR)</td>
<td>2.00 (1.00–2.00)</td>
</tr>
<tr>
<td></td>
<td>Spearman’s (ρ) P value</td>
<td>.06 P = .82</td>
</tr>
<tr>
<td>Advanced</td>
<td>Assessor score, median (IQR)</td>
<td>2.87 (2.31–3.25)</td>
</tr>
<tr>
<td>(n = 20)</td>
<td>Self-assessment, median (IQR)</td>
<td>3.0 (3.0–3.0)</td>
</tr>
<tr>
<td></td>
<td>Spearman’s (ρ) P value</td>
<td>−.11 P = .64</td>
</tr>
<tr>
<td>Experts</td>
<td>Assessor score, median (IQR)</td>
<td>3.13 (2.56–3.69)</td>
</tr>
<tr>
<td>(n = 20)</td>
<td>Self-assessment, median (IQR)</td>
<td>4.00 (3.00–4.00)</td>
</tr>
<tr>
<td></td>
<td>Spearman’s (ρ) P value</td>
<td>.12 P = .62</td>
</tr>
</tbody>
</table>

Spearman ρ correlation: <.30 was considered to be a weak correlation, between .30 and .50 was considered a moderate correlation, and >.50 was considered a strong correlation.
independent JAG-accredited assessors showed strong correlations for both polyp tasks, which implies that the assessor scores are in fact, accurate.

Finally, the process of self-assessment and independent assessment differed. Assessors watched video recordings retrospectively and therefore had time to scrutinize technique. Self-assessment scores were generated directly after the procedure. Martin et al.\textsuperscript{21} showed that the correlation between expert and self-assessment improves from .38 to .52 for communication skills if the participant reviews the video performance rather than from memory. This effect is diminished in more senior residents and experts.\textsuperscript{21} It would be interesting to see if this effect was more pronounced in the self-assessment of technical skills. Different

![Figure 3](image1)  
**Figure 3** Assessor 1 versus assessor 2 scores for the polyp A task (binned data according to scale [right of graph]).

![Figure 4](image2)  
**Figure 4** Assessor 1 versus assessor 2 scores for the polyp B task (binned data according to scale [right of graph]).
methods of conducting these appraisals should be explored in the future to ensure that self-assessment is utilized to its full potential in training and quality assurance.

A limitation of this trial may be that a simulator was used to replicate a real-life scenario. This was used to standardize the polypectomy task, which would be difficult to achieve in reality. This simulator has undergone rigorous validation studies previously but could never completely replicate real-life colonoscopic polypectomy. It may be interesting to assess the correlation between self-assessments and expert assessments using live polypectomy cases to see if the accuracy improves.

This research shows that there is a weak statistical correlation between self-assessment and independent expert assessment for simulated colonoscopic polypectomy. It remains to be seen if improvements in perceived performance may be achieved if participants self-assess video performance. Independent expert assessment seems to remain a strong way of reliably assessing performance in therapeutic colonoscopy.

References