Surg Endosc (2010) 24:1060–1065
DOI 10.1007/s00464-009-0726-6

Testing the construct validity of the Simbionix GI Mentor II virtual reality colonoscopy simulator metrics: module matters

Raad Fayez · Liane S. Feldman · Pepa Kaneva · Gerald M. Fried

Received: 20 April 2009 / Accepted: 6 October 2009 / Published online: 13 November 2009

Abstract
Background The use of simulation for competency assessment requires validation of the simulator's performance metrics. This study evaluated whether the Simbionix GI Mentor II virtual reality simulator metrics differentiate gastrointestinal endoscopists with varying clinical experience (known-groups construct validity).
Methods For this study, 20 subjects (medical and surgical) were classified into two groups based on self-reported clinical experience with colonoscopy: a novice group (<5 scope experiences, $n = 12$) and an experienced group (>50 scope experiences, $n = 8$). Three virtual colonoscopy simulation modules of increasing difficulty were used (modules I-1, II-2, and I-7). The data reported by the simulator after each module were compared using the Wilcoxon–Mann–Whitney test. Data are expressed as median and interquartile range (IQR). A p value less than 0.05 was considered statistically significant.
Results With module 1, only the time taken to reach the cecum was different between the groups: experienced group (1.6 min; IQR, 1.2–1.9 min) versus novice group (3.2 min; IQR, 2.4–4 min) ($p < 0.01$). With module 2, the two groups differed only in the time needed to reach the cecum (experienced group: 2.3 min; IQR, 1.6–2.3 min vs novice group: 3.3 min; IQR, 2.3–4.2 min; $p = 0.03$) and overall efficiency (experienced group: 94%; IQR, 94–96% vs novice group: 88%, IQR, 69–92%) ($p < 0.01$). In contrast, with the module 3 (the most difficult), performance differed between the groups for most of the parameters. The experienced group reached the cecum faster (5.7 min; IQR, 3.6–6.6 min vs. 14 min; IQR, 9–16 min; $p < 0.01$) and had fewer occasions of lost view (0.5; IQR, 0–1 vs. 2; IQR, 2–3; $p < 0.01$), fewer episodes of excessive pressure (2; IQR, 1–2 vs. 4.5; IQR, 2.5–6; $p < 0.01$), and greater overall efficiency (87%; IQR, 82–89% vs. 29%; IQR, 23–55%; $p < 0.01$). There were no differences in the percentage of time the patient was in pain or in the total time the colon was looped. The experienced group saw slightly less of the mucosa (91%; IQR, 89–92% vs 94%; IQR, 93–95%; $p = 0.01$).
Conclusion The GI Mentor II metrics differentiated novice colonoscopists from those with more clinical experience, but primarily when used to evaluate the more complex scenarios. In setting performance benchmarks, the case scenario must be taken into account.

Keywords Colonoscopy · Flexible endoscopy · Simulator · Training · Validation · Virtual reality

Presented at the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) meeting at Phoenix, Arizona, April 2009.
This study was funded by an unrestricted educational grant from Covidien Canada. The Steinberg-Bernstein Centre for Minimally Invasive Surgery is supported by an unrestricted educational grant from Covidien Canada.

R. Fayez · L. S. Feldman · P. Kaneva · G. M. Fried
Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, McGill University, Montreal, Quebec, Canada

G. M. Fried (✉)
Montreal General Hospital, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4, Canada
e-mail: gerald.fried@mcgill.ca

Introduction

Flexible endoscopy is a vital component in the diagnosis and treatment of digestive disease. However, endoscopy is relatively invasive and may be associated with discomfort

and adverse events, particularly during the learning curve [1]. Demands for procedural efficiency combined with resident work hour restrictions, the increased focus on medical errors, and the ethics of "practicing" on patients all have stimulated interest in the role of simulation as an adjunct to traditional procedural skills training. The educational advantage of simulation is that it enables the development of reproducible curricula that allow for safe repetition of standardized cases in a nonstressful environment. In addition to this educational role, simulation allows for measurement of performance using objective metrics, with a potential role in competency assessment and credentialing [2].

A number of virtual reality flexible endoscopy simulators are available currently [3], in which high-fidelity reproductions of flexible endoscopy are presented through a number of simulated cases varying in difficulty. At the end of each simulated case, the user is provided with a long list of metrics evaluating his or her performance. Evidence shows that simulation-based training in colonoscopy can be used both as a beneficial educational tool during the initial stages of learning [4]. However, there is also interest in the use of simulation metrics for the assessment of competency. For these evaluations to be used in high-stakes competency assessment, supporting evidence for their validity must be provided (i.e., evidence that the metrics measure what they purport to measure, namely, proficiency in endoscopy). Because no current "gold standard" method exists for measuring clinical endoscopic performance (i.e., criterion validity), evidence for construct validity may be sought by comparing simulator performance of subjects with varying degrees of clinical endoscopic experience.

This study aimed to evaluate whether the Simbionix GI Mentor II (Simbionix USA Corp., Cleveland, OH, USA) virtual reality colonoscopy simulator metrics differentiate endoscopists with varying degrees of clinical experience (known-groups construct validity). Although previous studies also have assessed this [5–7], we particularly sought to determine the impact of the chosen case module on the measured performance and to identify which of the multiple reported metrics had the greatest discriminatory value.

Methods

Simulator

The Simbionix GI Mentor II flexible endoscopy simulator was used to assess the performance of colonoscopy [8]. With the Simbionix GI Mentor II, the procedures are performed on a mannequin using an adapted Pentax endoscope with controls, including suction and inflation buttons, analogous to a real endoscope used in a clinical sitting (Fig. 1). Sensors transmit information about the location of the scope as it is advanced in the mannequin. Visual cues (changes in the colonic wall seen on the monitor) and a force-feedback sensation simulate tissue resistance encountered by the scope.

Fig. 1 The Simbionix GI Mentor II flexible endoscopy simulator consisting of a mannequin, monitor, and flexible scope

Two colonoscopy modules exist, each with 10 virtual cases. The cases differ in complexity based on the degree of intestinal looping and pathologies encountered. Numerous parameters are reported at the end of each module. These include the time taken to reach the cecum (minutes), the number of times the view of the lumen is lost, the number of times excessive pressure occurs, the percentage of mucosa visualized, the percentage of time the view is clear, the percentage of time a patient is in pain, the total time colon is looped (seconds), and the percentage efficiency of the screening.

Participants and procedures

For this study, 20 subjects were classified into two groups based on their self-reported clinical experience with lower gastrointestinal (GI) endoscopy (12 novices and 8 experienced subjects). The experienced group included five senior surgical residents and gastroenterology fellows with a mean of 80 previous colonoscopies, and three attending staff in surgery and gastroenterology with more than 1,000 previous colonoscopies each. With the exception of one subject who preformed two colonoscopies, the junior surgical residents and gastroenterology fellows in the novice group had no prior colonoscopy experience. Three subjects in the experienced group and one in the novice group reported that they had tried the simulator previously. The study was approved by the Research Ethics Board of McGill University (GEN# 08-050), and the participants gave informed consent. It was conducted at the McGill University Simulation Centre.

After identical initial instructions, three virtual colonoscopy simulations of increasing difficulty were performed by

each subject. Module 1 (I-1) had an average colon with no pathologic findings, whereas module 2 (II-1) had an average colon with a pedunculated polyp in the sigmoid colon; and module 3 (I-7) had a long sigmoid with a long and twisted hepatic flexure, a sessile polyp in the transverse colon, and a pedunculated polyp in the descending colon (Fig. 2).

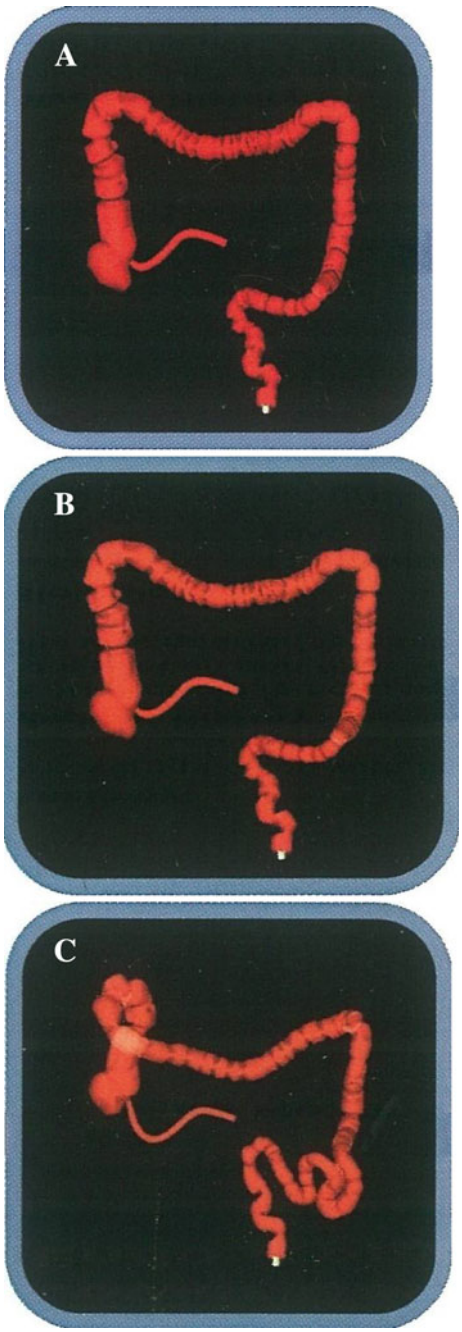

Fig. 2 Modules used in the study. **A** Module 1 (I-1) has an average colon with no pathologic findings. **B** Module 2 (II-1) has an average colon with a pedunculated polyp in the sigmoid colon. **C** Module 1 (I-7) has a long sigmoid with a long and twisted hepatic flexure, a sessile polyp in the transverse colon, and a pedunculated polyp in the descending colon

The subjects were told to perform polypectomy as indicated. No practice sessions were allowed. All aspects of the procedure, including insertion, navigation, and advancing of the scope were done by the operator. The participants were allowed to examine the report generated at the completion of each module.

Statistical analysis

The data were analyzed using SAS version 9.1.3 (SAS Institute Inc., Cary, NC, USA). Differences in the performance parameters generated after each module were compared between the novice and experienced groups using the Wilcoxon–Mann–Whitney test. Data are expressed as median and interquartile range (IQR). *P* value less than 0.05 was considered statistically significant.

Results

On the first and easiest simulation (module 1, case 1), only the time taken to reach the cecum was different between the groups (experienced group: 1.6 min; IQR, 1.2–1.9 min vs. novice group: 3.2 min; IQR, 2.4–4 min; $p = 0.001$). There were no significant differences between the groups for the other reported parameters (Table 1). In the second case (module 2, case 1), only time and efficiency of screening were significantly different between the groups (Table 2). In contrast, in the third and most difficult simulation (module 1, case 7), the groups differed in most performance parameters. The experienced group reached the cecum faster (5.7 min; IQR, 3.6–6.6 min vs. 14 min; IQR, 9–16 min; $p<0.01$) and had a greater proportion of time with a clear view (92%; IQR, 91–94% vs. 88%; IQR, 87–93%; $p = 0.05$), fewer occasions of lost view (0.5; IQR, 0–1 vs. 2; IQR, 2–3; $p < 0.01$), fewer episodes of excessive pressure (2; IQR, 1–2 vs. 4.5; IQR, 2.5–6; $p < 0.01$), and greater overall efficiency (87%; IQR, 82–89% vs. 29%; IQR, 23–55%; $p < 0.01$). There were no differences in the percentages of time the patient was in pain or the total time the colon was looped. The experienced group saw slightly less of the mucosa (91%; IQR, 89–92% vs. 94%; IQR, 93–95%; $p = 0.01$) (Table 3). The last two modules also involved identification and removal of polyps. This was successfully completed by all the participants with both modules.

Discussion

Procedural simulation can be used for the purposes of both education and objective assessment. The educational value of virtual reality colonoscopy and sigmoidoscopy

Table 1 Comparison of performance parameters of novices (<5 scope experiences) and more experienced endoscopists (>50 scope experiences) in the first simulation (module 1, case 1)

	Time (min)	Lost lumen (*n*)	Excessive pressure (*n*)	Mucosa seen (%)	Clear view (%)	Pain (%)	Looping (s)	Efficiency (%)
Novice	3.2 (2.4–4)	0 (0–0)	1 (0.5–1.5)	87 (85–90)	95 (94–97)	0 (0–3.5)	0 (0–26)	91 (71–92)
Experienced	1.6 (1.2–2)	0 (0–0)	0.5(0–1)	85 (81–87)	96 (94–98)	2 (0.5–4)	11 (3–24)	94 (87–95)
p Value	<0.01	0.26	0.14	0.13	0.61	0.30	0.38	0.08

Values are expressed as median (interquartile range)

Table 2 Comparison of performance parameters of novices (<5 scope experiences) and more experienced endoscopists (>50 scope experiences) in the second simulation (module 2, case 1)

	Time (min)	Lost lumen (*n*)	Excessive pressure (*n*)	Mucosa seen (%)	Clear view (%)	Pain (%)	Looping (s)	Efficiency (%)
Novice	3.3 (2.3–4.2)	0 (0–0)	0 (0–1)	91 (89–91)	97 (96–98)	0 (0–0)	0.5 (0–3)	88 (69–92)
Experienced	2.3 (1.6–2.3)	0 (0–0)	0 (0–0)	87 (86–89)	97 (96–98)	0 (0–0.5)	3.5 (0–8.5)	94 (94–96)
p Value	0.03	1	0.42	0.07	0.11	0.90	0.64	<0.01

Values are expressed as median (interquartile range)

Table 3 Comparison of performance parameters of novices (<5 scope experiences) and more experienced endoscopists (>50 scope experiences) in the third simulation (module 1, case 7)

	Time (min)	Lost lumen (*n*)	Excessive pressure (*n*)	Mucosa seen (%)	Clear view (%)	Pain (%)	Looping (s)	Efficiency (%)
Novice	14 (9–16)	2 (2–3)	4.5 (2.5–6)	94 (93–95)	88 (87–93)	0 (0–0.5)	4 (0–22)	29 (23–55)
Experienced	5.7 (3.6–6.6)	0.5 (0–1)	2 (1–2)	91 (89–92)	92 (91–94)	0 (0–0)	4 (0–19)	87 (82–89)
p Value	<0.01	<0.01	<0.01	0.01	0.05	0.58	0.93	<0.01

Values are expressed as median (interquartile range)

simulation has been supported in systematic reviews of the evidence, with benefits seen primarily in the earliest phase of training [4, 9, 10]. However, for simulation to be used for competency assessment, a high-stakes examination, strong evidence for the validity of the performance metrics, is required. Ultimately, performance in the simulator should correlate with clinical colonoscopy performance and demonstrate ability to predict such performance (predictive validity). As a first step, however, the metrics should be able to differentiate between subjects with differing clinical experience levels (construct validity).

The current study adds to evidence supporting the construct validity of the performance metrics reported for the GI Mentor II after a simulated colonoscopy scenario. Compared with novices, the more clinically experienced group was faster, had fewer incidences of excessive pressure, and were more efficient in screening the mucosa. However, these differences were noted only in the most difficult case scenario, with relatively few differences seen with the two easier cases. This finding has implications for the potential use of simulation for competency assessment. If the module is not sufficiently difficult, the metrics will not differentiate users with differing clinical competencies.

Although previous studies have provided evidence for the construct validity of the GI Mentor II colonoscopy performance metrics, the differences between groups often is relatively small. Felsher et al. [5] recruited 75 subjects at two Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) events. They were classified into two groups based on their training level and clinical endoscopy case volume. At the SAGES meeting, 37 participants were tested on module 1, case 5. There were statistically significant differences between the groups. The more experienced group saw slightly more mucosa, were more likely to perform polypectomy, and were somewhat more efficient ($32 \pm 8\%$ vs. $26 \pm 7\%$). There were no significant differences in time (4.8 ± 1.2 min vs. 5.6 ± 1.4 min) or pain. A cohort of 38 subjects was tested in a resident course on module 2, case 1. The experienced subjects were faster (2.9 ± 1.3 min vs. 4.4 ± 1.8 min), more likely to perform polypectomy, and somewhat more efficient ($53 \pm 8\%$ vs. $37 \pm 7\%$). There were no differences in mucosa seen, time with clear view, or pain. The influence of the module on performance was demonstrated by the fact that the novice group's efficiency with module 2, case 1 was higher than the experienced group's efficiency with module 1, case 5.

Grantcharov et al. [6] classified 28 subjects into three groups based on previous endoscopy experience as follows: experienced subjects (>200 procedures), residents (<50 procedures), and medical students (no procedures). The

module and case assessed were not stated. The experienced group performed better than the residents and the students, but there were no differences between the medical students and the residents. The medical students reached the cecum in an average of about 5 min and had more than 80% screening efficiency. This suggests that the case selected was quite easy because the screening efficiency for the students with no clinical experience was higher than that reported for the residents and surgeons in the Felsher et al. [5] study.

The largest study to date evaluating the construct validity of the GI Mentor II colonoscopy metrics enrolled 105 subjects at the Dutch Society of Gastroenterology and a hospital site [7]. The subjects were classified into four groups, namely, residents (no scope experiences), intermediate subjects (<200 colonoscopies), experienced subjects (200–1,000 colonoscopies), and experts (>1,000 colonoscopies). They were assessed on module 1, cases 1 and 3. With both simulations, the differences between the four groups were due to the novices performing worse than the other groups, with no differences seen between any of the other three groups. In other words, the metrics could differentiate between a group with no previous scope experience and one with more than 200 scope experiences, but there were no differences detected in the three groups with more than 200 scope experiences. This difference was more pronounced for case 3 than for case 1. According to a questionnaire, the experts thought the simulator could be a teaching tool for novices but did not think the simulation was "suitable for certification of trained endoscopists."

In an effort to find an expert benchmark for performance with the GI Mentor II, Phitayakorn et al. [11] assessed 23 expert colonoscopists (all with >1,000 scope experiences) on module 1, case 5 at a SAGES conference. The mean time taken to reach the cecum was 6.5 ± 4.3 min (range, 2.8–20.2 min). The mean overall efficiency was 70 ± 24% (range, 20–94%). They noted that even among expert colonoscopists, considerable variability existed in simulator performance.

The number of performance metrics generated after each case also complicated benchmark setting. We found that many of the metrics were not very different between the two groups, except for the most difficult case. However, time and percentage of mucosa seen were significantly different, or approached significance, in all three cases.

The GI Mentor estimates the percentage of mucosa examined by dividing the colon wall into 2,000 equal areas along its surface. Each area is considered screened only if it is visible on the screen for more than half a second. Time and percentage of mucosa screened are used to define the efficiency of screening, which may be a good candidate as a benchmark setting for these simulations.

Events such as excessive pressure, lost lumen view and looping were relatively rare even in the most difficult case and may be best used as "errors" in setting benchmarks. For example, proficiency levels may be defined as attaining a certain percentage of efficiency with a maximum "allowable" number of errors. This approach requires further investigation with a larger number of expert subjects.

A limitation of the current study was the relatively small sample size. The subjects also were quite heterogeneous including gastroenterology fellows, surgical residents, and attending staff in gastroenterology and surgical endoscopy. This may have affected the ability of the metrics to detect subtle competency differences in the first two scenarios, although large differences were not apparent.

We originally chose a cutoff of 50 previous endoscopies to define the novice group because this is the number required by the American Council for Graduate Medical Education (ACGME) for graduating surgical residents [12]. However, the novice group in fact had virtually no previous clinical experience, which should have maximized the likelihood that the metrics would differentiate between the two groups in the study. The experienced group had a wide range of previous endoscopies, including that of three experts as well as gastroenterology fellows and senior surgical residents. However, these finer gradations in experience were not detected by the performance metrics when used in previous validation studies [7].

In conclusion, evidence in this study supports the construct validity of the performance metrics for the GI Mentor II colonoscopy simulator. The percentage efficiency parameter may be a good candidate for further investigation to define proficiency targets. The ability of the metrics to differentiate between experience levels in colonoscopy may be amplified when more difficult case scenarios are chosen, and future validation studies should take this factor into account.

Disclosures Raad Fayez, Liane S. Feldman, Pepa Kaneva, and Gerald M. Fried have no conflicts of interest of financial ties to disclose.

References

1. Bini EJ, Firoozi B, Choung RJ, Ali EM, Osman M, Weinshel EH (2003) Systematic evaluation of complications related to endoscopy in a training setting (SECRETS): a prospective 30-day outcomes study. Gastrointest Endosc 57:8–16
2. Fried GM, Feldman LS (2008) Objective assessment of technical performance. World J Surg 32:156–160
3. Greenwald D, Cohen J (2006) Evolution of endoscopy simulators and their application. Gastrointest Endosc Clin North Am 16: 389–406

4. Sturm LP, Windsor JA, Cosman PH, Cregan P, Hewett PJ, Maddern GJ (2008) A systematic review of skills transfer after surgical simulation training. Ann Surg 248:166–179
5. Felsher JJ, Olesevich M, Farres H, Rosen M, Fanning A, Dunkin BJ, Marks JM (2005) Validation of a flexible endoscopy simulator. Am J Surg 189:497–500
6. Grantcharov TP, Carstensen L, Schulze S (2005) Objective assessment of gastrointestinal endoscopy skills using a virtual reality simulator. JSLS 9:130–133
7. Koch AD, Buzink SN, Heemskerk J, Botden SMBI, Veenendaal R, Jakimowicz JJ, Schoon EJ (2007) Expert and construct validity of the Simbionix GI Mentor II endoscopy simulator for colonoscopy. Surg Endosc 22:158–162
8. Bar-Meir S (2006) Simbionix simulator. Gastrointest Endosc Clin North Am 16:471–478
9. Gerson LB (2006) Evidence-based assessment of endoscopic simulators for training. Gastrointest Endosc Clin North Am 16:489–509
10. Cohen J, Cohen SA, Vora KC, Xue X, Burdick JS, Bank S, Bini EJ, Bodenheimer H, Cerulli M, Gerdes H et al (2006) Multicenter, randomized, controlled trial of virtual-reality simulator training in acquisition of competency in colonoscopy. Gastrointest Endosc 64:361–368
11. Phitayakorn R, Marks JM, Reynolds HL, Delaney CP (2009) Expert benchmark for the GI mentor II. Surg Endosc 23:611–614
12. ACGME memorandum (2006) Changes in minimum requirements for laparoscopy and endoscopy. Retrieved 9 March 2009 at www.acgme.org/acWebsite/RRC_440/440_minReqLaparoscopy.asp