Validity evidence for AI-generated complexity scores of laparoscopic cholecystectomy

Dylan Grisell MD[1,3], Saba Balvardi MD[1,3], Melina Vassiliou MD MEd[1,3], Mohammed Al Mahroos MD[1], Pepa Kaneva MSc[1,2] Julio F. Fiore Jr PhD[1,2,3], Liane S Feldman MD[1,2,3]

[1] *Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University, Montreal, QC, Canada;*
[2] *Centre for Outcomes Research and Evaluation (CORE),Research Institute of the McGill University Health Centre, Montreal, QC, Canada;* [3] *Division of General Surgery, Department of Surgery, McGill University, Montreal, QC, Canada*

Key Message

Artificial Intelligence can produce valid assessments of case complexity for videos of laparoscopic cholecystectomy

Introduction

- Video review improves laparoscopic skill acquisition[1], but the time required for expert analysis is a barrier to implementation in surgical training programs[2]
- Artificial Intelligence (AI) programs may reduce this time burden by providing automated evaluations that should account for case complexity

Objective – To contribute validity evidence for AI-generated complexity scores of laparoscopic cholecystectomy (LC) cases performed by general surgical trainees

Methods

Collection & Storage: Supervised LC cases were recorded and uploaded to a secure web-accessible platform (Theator, Palo Alto) in a previous trial[3]

Clinical Measures: Procedure time, diagnosis, and priority

Case Complexity Scores:

- Intraoperative expert assessment by Visual Analogue Scale (VAS, 1-10)
- Blinded expert assessment by VAS after full video review (Blind-VAS, 1-10)
- AI-generated case complexity based on the Parkland Grading Scale[4] (AI-PGS, 1-5)
  - 1 = Normal gallbladder → 5 = Perforation/necrosis/obscuring adhesions

Validity Hypotheses: If AI-PGS is a valid measure of complexity, it will correlate with:

- Primary: **(H1)** VAS and **(H2)** Blind-VAS
- Secondary: **(H3)** Procedure time, **(H4)** Diagnosis, and **(H5)** Priority

Data Analysis: COSMIN[5] checklist used to formulate hypotheses & report results

- **H1-3**: Spearman's rank correlation
  - Confirmed if correlation coefficient is at least moderate (Rho ≥ 0.4)
- **H4-5**: Mann Whitney test for between-group median differences
  - Confirmed if between-group median difference (BGMD) is ≥1

Results

- 108 LC cases; 31 trainees; 16 supervising surgeons; 3 blinded expert reviewers

Table 1: Clinical measures & median AI-PGS for subgroups

Diagnosis	**N (%)**	**Median AI-PGS**
Biliary Colic/Polyp	51 (47%)	2
Cholecystitis/CBD Stone/Pancreatitis	57 (53%)	3
Priority	**N (%)**	**Median AI-PGS**
Emergency	67 (62%)	3
Elective	41 (38%)	3
Procedure time	**Mean (SD)**	
Time (min)	87.6 (30.1)	

Table 2: Case complexity scores

Case complexity score (range)	**Mean (SD)**	**Median**	**Mode**
VAS (1-10)	4.00 (2.69)	3	3
Blind-VAS (1-10)	3.89 (2.64)	3	2
AI-PGS (1-5)	2.66 (1.09)	3	3

Fig 1: HI – AI-generated complexity (AI-PGS) vs. Intraoperative complexity (VAS)

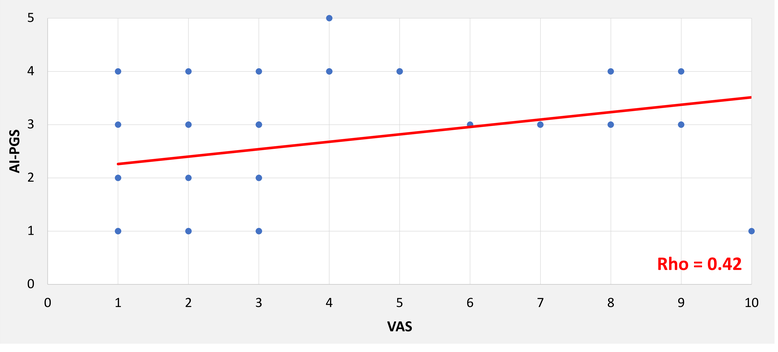


Fig 2: H2 – AI-generated complexity vs. Blinded expert assessment (Blind-VAS)

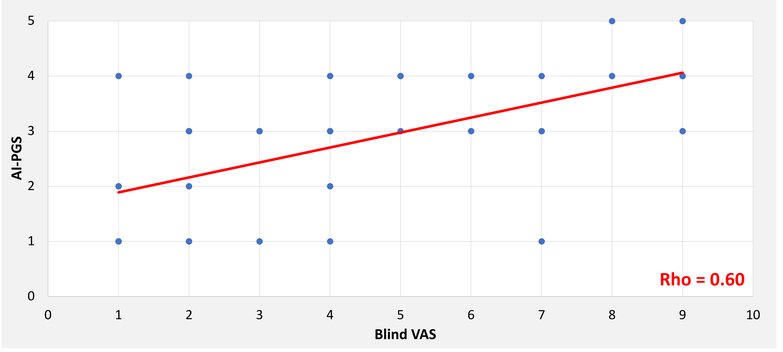


Table 3: Validity hypothesis testing

Hypothesis	**Rho (p)**	**BGMD (p)**	**Confirmed?**
H1	0.42 (<0.01)	-	Yes
H2	0.60 (<0.01)	-	Yes
H3	0.27 (<0.01)	-	No
H4	-	1 (<0.01)	Yes
H5	-	0 (<0.02)	No

Conclusion – This study contributes evidence supporting the validity of AI-generated complexity scores for video recordings of laparoscopic cholecystectomy

References

1. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP. Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Ann Surg*. 2015;262(2):205-212
2. Snyder RA, Tarpley MJ, Tarpley JL, et al. Teaching in the Operating Room: Results of a National Survey. *J Surg Ed*. 2012;69(5):643-9.
3. Balvardi S, Semsar-Kazerooni K, Kaneva P, et al. Validity of video-based general and procedure-specific self-assessment tools for surgical trainees in laparoscopic cholecystectomy. *Surg Endosc*. 2023;37(3):2281-9.
4. Madni TD, Leshikar DE, Minshalll CT, et al. The Parkland grading scale for cholecystitis. *Am J Surg*. 2018; 215(4): 625-630.
5. Mokkink L, Terwee C, Patrick D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Quality of Life Research 2010;19:539-549.