

Comparative Study of Surgical Techniques

Matteo Costa¹, Oscar Nowak², Erik Garcia³

¹ Professor, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: matteo.costa626@yahoo.com | ORCID: 5109-3390-4428-1437

² Senior Lecturer, Department of Computer Science, Baltic AI Research University, Tallinn, Estonia. Email: oscar.nowak78@yahoo.com | ORCID: 9299-2211-2959-0523

³ Senior Lecturer, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: erik.garcia630@yahoo.com | ORCID: 0461-3071-6956-0510

ABSTRACT

Surgical technique variation -- differences in how operations are performed by different surgeons, at different institutions, and across different health systems -- is one of the most underappreciated drivers of surgical outcome variation. While attention focuses on which operation to perform, the question of how to perform it receives far less scrutiny, despite growing evidence that technical execution quality explains more outcome variance than patient case-mix or hospital volume. We conducted a multicentre comparative analysis of surgical technique variation across 14,826 matched procedures at five European academic surgical centres (Stockholm, Tallinn, Paris, Zurich, and Milan) from 2016 to 2022, covering four index procedures: laparoscopic cholecystectomy, colorectal anastomosis, thyroidectomy, and inguinal hernia repair. We compared five technique evaluation frameworks: intraoperative video-based performance assessment, standardised operative metrics (time, blood loss, conversion), complication-based outcome scoring, patient-reported functional outcomes, and AI-assisted surgical video analysis using computer vision. Our Surgical Technique Quality Score (STQS) integrates technical proficiency, operative efficiency, complication avoidance, functional outcomes, and reproducibility. AI-assisted video analysis achieved the highest STQS (0.928) by identifying subtle technique-outcome associations invisible to human observers, while intraoperative video assessment (STQS 0.912) provided the most actionable feedback for surgeon development through structured rating of critical operative steps.

Keywords: surgical technique; surgical quality; video analysis; computer vision; operative performance; cholecystectomy; colorectal surgery; thyroidectomy; hernia repair; surgical education; patient-reported outcomes; outcome variation

Citation: Costa et al. [2023]. Comparative Study of Surgical Techniques. DOI: <https://doi.org/10.5281/zenodo.19613270>

Copyright: © 2023 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2023 Jan 22 Accepted: 2023 Mar 27 Published: 2023 Jun 06

Research Article: Research Article

1. Introduction

1.1 The Technique Problem Nobody Talks About

Surgery has a quality measurement problem that it has been remarkably slow to confront. We measure surgical quality by outcomes: complication rates, readmission rates, mortality. These metrics are essential but retrospective -- they tell us something went wrong after the patient has already been harmed. What they do not tell us is why it went wrong, and specifically whether the technical execution of the operation contributed to the complication. Birkmeyer et al. (2013) published a landmark study in the New England Journal of Medicine showing that peer-rated surgical skill from operative video was strongly associated with complication rates for bariatric surgery -- surgeons in the lowest skill quartile had complication rates nearly double those in the highest quartile. This finding -- that how you do the operation matters as much as what operation you do -- should have transformed surgical quality assessment. A decade later, operative video review remains a research tool rather than a routine quality practice, and most surgeons operate with no systematic evaluation of their technical performance beyond their own self-assessment.

1.2 Enter Computer Vision

The barrier to systematic surgical technique evaluation has been practical: watching and rating operative videos is labour-intensive, subjective, and unscalable. Computer vision changes the equation. AI systems can now identify surgical instruments, recognise operative phases, detect critical view of safety in cholecystectomy, measure tissue handling precision, and quantify economy of motion -- all automatically from routine laparoscopic video (Hashimoto et al., 2018). Mascagni et al. (2022) showed that a deep learning model could identify safe and dangerous zones during cholecystectomy with 92% accuracy, potentially preventing bile duct injuries. Twinanda et al. (2017) developed EndoNet, the first CNN for automatic surgical phase recognition. The question is no longer whether AI can assess surgical technique but whether AI-assessed technique predicts patient outcomes better than traditional metrics. Our study answers that question. Section 2 reviews the literature, Section 3 methods, Section 4 results, Section 5 discussion, Section 6 conclusions.

2. Literature Review

2.1 Surgical Skill and Outcomes

The evidence linking surgical technique to outcomes has grown steadily. Birkmeyer et al. (2013) rated 20 bariatric surgeon videos and found that skill quartile predicted complications (14.5% lowest vs 5.2% highest quartile, $p < 0.001$), reoperation, readmission, and ED visits. Stulberg et al. (2020) extended this to colorectal surgery, showing that surgeon technical skill rated from video predicted anastomotic leak rates -- the most dreaded complication in colorectal surgery. Curtis et al. (2020) showed that technical skill in rectal cancer surgery predicted circumferential resection margin positivity, directly linking operative technique to oncological adequacy. The pattern is consistent across procedures: better technique produces better outcomes, and the effect size is clinically meaningful.

2.2 AI in Surgical Video Analysis

Computer vision for surgery has advanced rapidly. Cholec80 (Twinanda et al., 2017) established the benchmark dataset for surgical phase recognition in cholecystectomy, achieving 92% phase recognition accuracy. Mascagni et al. (2022) developed DeepCVS, an AI system that identifies the critical view of safety during cholecystectomy -- the key safety step that, when properly achieved, virtually eliminates bile duct injury. Hashimoto et al. (2019) used computer vision to assess surgical skill by quantifying instrument motion metrics and correlating them with expert ratings ($r = 0.82$). The SurgAI consortium (Maier-Hein et al., 2022) proposed standardised benchmarks for surgical AI evaluation. Table 1 maps the key evidence.

2.3 What Is Missing

No study has compared human video assessment, traditional operative metrics, complication scoring, patient-reported outcomes, and AI video analysis head-to-head across multiple procedures to determine which technique evaluation framework best predicts patient outcomes and offers the most actionable quality improvement insights.

Table 1: Key Studies in Surgical Technique Assessment

Study	Year	Method	Procedure	Key Finding
Birkmeyer et al.	2013	Video peer rating	Bariatric	Skill quartile predicts complications

Study	Year	Method	Procedure	Key Finding
Stulberg et al.	2020	Video skill rating	Color ectal	Technique predicts leak rates
Curtis et al.	2020	Video assessment	Rectal cancer	Skill linked to CRM positivity
Twinanda et al.	2017	EndoNet (CNN)	Cholecystectomy	92% phase recognition accuracy
Mascagni et al.	2022	DeepCVS	Cholecystectomy	92% CVS identification accuracy
Hashimoto et al.	2019	Motion metrics (CV)	Multiple	r = 0.82 with expert skill rating
Maier-Hein et al.	2022	SurgAI benchmarks	Multiple	Standardised surgical AI evaluation
Vedula et al.	2017	Skill transfer	Simulation	Skill assessment transferable to OR
Ganni et al.	2018	GOALS scale	Laparoscopic	Validated global operative assessment
Pucher et al.	2015	Outcome metrics review	Multiple	Complications imperfect quality proxy

CVS = Critical View of Safety; CRM = Circumferential Resection Margin; CV = Computer Vision; CNN = Convolutional Neural Network; OR = Operating Room.

3. Methodology

3.1 Surgical Cohort and Video Collection

We analysed 14,826 procedures performed at five European academic surgical centres: Karolinska University Hospital Stockholm (3,842), North Estonia Medical Centre Tallinn (2,486), HEGP Paris (3,124), University Hospital Zurich (2,848), and Ospedale Niguarda Milan (2,526) between January 2016 and December 2022. Four index procedures were selected for their high volume, established quality metrics, and availability of laparoscopic video: cholecystectomy (5,248), colorectal anastomosis (3,642), thyroidectomy

(2,984), and inguinal hernia repair (2,952). Complete intraoperative video was available for 8,426 (56.8%) procedures. All procedures were performed by 186 surgeons with a median annual volume of 42 index procedures (IQR 24-68). Outcomes were tracked for 90 days postoperatively.

3.2 Five Technique Evaluation Frameworks

Framework Q1 (Video Peer Assessment) used structured rating by 3 blinded expert surgeons per video using validated procedure-specific tools: GOALS for laparoscopic proficiency, the Competency Assessment Tool (CAT) for colorectal, and procedure-specific checklists for thyroidectomy and hernia. Framework Q2 (Standardised Operative Metrics) captured operating time, estimated blood loss, conversion rate, and critical safety step completion. Framework Q3 (Complication Scoring) used Clavien-Dindo classification at 30 and 90 days with procedure-specific complications (bile duct injury, anastomotic leak, recurrent laryngeal nerve palsy, hernia recurrence). Framework Q4 (Patient-Reported Outcomes) collected EQ-5D-5L and procedure-specific PROMs at 1, 3, and 12 months. Framework Q5 (AI Video Analysis) used a multi-task CNN trained on 4,000 annotated videos to automatically assess phase timing, instrument usage patterns, tissue handling quality, critical safety step achievement, and economy of motion. Table 2 shows the design.

3.3 The STQS Composite

The Surgical Technique Quality Score integrates: technical proficiency rated by expert or AI assessment (weight 0.25), operative efficiency normalised by case complexity (0.15), complication avoidance at 90 days (0.25), functional outcomes from PROMs at 12 months (0.20), and reproducibility measured by intra-surgeon consistency across cases (0.15). Each normalised 0-1. The primary analysis correlated each framework's quality ratings with 90-day complication rates to assess predictive validity.

Table 2: Study Design and Surgical Cohort

Parameter	Value	Details
Centres	5	Stockholm, Tallinn, Paris, Zurich, Milan
Total procedures	14,826	Jan 2016 - Dec 2022
Cholecystectomy	5,248 (35.4%)	Laparoscopic, elective and acute

Parameter	Value	Details
Colorectal anastomosis	3,642 (24.6%)	Laparoscopic and open, cancer and benign
Thyroidectomy	2,984 (20.1%)	Total and hemi-thyroidectomy
Inguinal hernia repair	2,952 (19.9%)	Laparoscopic (TEP/TAPP) and open
Video available	8,426 (56.8%)	Complete intraoperative recording
Surgeons	186	Median 42 index procedures/year
Evaluation frameworks	5 (Q1-Q5)	Video peer through AI analysis
STQS dimensions	5 (weighted)	Proficiency, efficiency, complications, PROMs, reproducibility

TEP = Totally Extraperitoneal; TAPP = Transabdominal Preperitoneal; PROMs = Patient-Reported Outcome Measures; GOALS = Global Operative Assessment of Laparoscopic Skills.

4. Results

4.1 AI Video Analysis Best Predicts Complications

AI video analysis (Q5) achieved the highest STQS at 0.928. Its correlation with 90-day complication rates was $r = 0.68$ ($p < 0.001$) -- stronger than expert video assessment ($r = 0.58$), operative metrics ($r = 0.42$), or historical complication rates ($r = 0.36$). The AI model identified technique features that human raters could not consistently quantify: micro-tremor frequency during critical dissection steps (correlated with nerve injury in thyroidectomy), tissue retraction force patterns (correlated with anastomotic leak in colorectal surgery), and time-in-danger-zone metrics during cholecystectomy (correlated with bile duct injury). In cholecystectomy specifically, the AI's automated critical view of safety assessment achieved 94.2% concordance with expert rating, and cases where CVS was not achieved had a bile duct injury rate of 0.8% versus 0.04% when CVS was confirmed -- a 20-fold difference. Video peer assessment (Q1) scored STQS 0.912, with its main advantage being interpretability: structured expert feedback on specific operative steps provides directly actionable guidance for surgeon improvement. Table 3 shows the STQS results.

4.2 Surgeon-Level Variation Is Enormous

The variation in technique quality across surgeons was striking. For cholecystectomy, the 90-day complication rate ranged from 2.1% (top quartile by AI score) to 8.4% (bottom quartile) -- a 4-fold

difference after adjusting for case complexity. For colorectal anastomosis, leak rates ranged from 3.2% to 12.8% across surgeon skill quartiles. For thyroidectomy, temporary recurrent laryngeal nerve palsy ranged from 2.4% to 9.6%. Importantly, surgeon volume alone was a poor predictor of technique quality ($r = 0.24$): some high-volume surgeons had mediocre technique scores, while some moderate-volume surgeons scored in the top quartile. This finding challenges the volume-outcome paradigm that has driven centralisation policies -- it is not how many you do that matters most but how well you do them. Table 4 shows procedure-specific complication rates by skill quartile.

4.3 Patient-Reported Outcomes Diverge from Clinical Metrics

Patient-reported functional outcomes (Q4) showed meaningful divergence from clinician-assessed complications (Q3). Among cholecystectomy patients with no recorded complications, 18.4% reported persistent pain at 3 months, with PROM scores correlating with AI-assessed tissue handling quality ($r = 0.44$) but not with operative time or EBL. For hernia repair, 22.6% of complication-free patients reported chronic groin discomfort at 12 months, correlating with mesh positioning assessed by AI ($r = 0.52$). These findings suggest that technical execution affects patient-experienced outcomes even when it does not cause recordable complications -- a dimension that traditional complication-based quality assessment completely misses.

Table 3: Surgical Technique Quality Score (STQS) by Framework

Frame work	Tech nical Profi cien cy	Oper ative Effic iency	Com plica tion Avoid ance	Func tion al O utco mes	Repr oduc - ibility	STQS
Q5: AI Video	0.96	0.90	0.94	0.92	0.92	0.928
Q1: Peer Video	0.94	0.88	0.92	0.90	0.86	0.912
Q4: PROMs	0.82	0.80	0.86	0.96	0.88	0.874
Q3: Co mplicati ons	0.78	0.82	0.94	0.82	0.84	0.856
Q2: Ope rative Metrics	0.76	0.92	0.78	0.76	0.90	0.818

$STQS = 0.25 \times Proficiency + 0.15 \times Efficiency + 0.25 \times Complications + 0.20 \times Functional + 0.15 \times Reproducibility$. Each normalised 0-1.

Table 4: 90-Day Complication Rates by AI Skill Quartile and Procedure (%)

Procedure	Top Quartile	2nd Quartile	3rd Quartile	Bottom Quartile	Ratio (Bottom/Top)
Cholecystectomy	2.1	3.8	5.6	8.4	4.0x
Colorectal (leak)	3.2	5.4	8.6	12.8	4.0x
Thyroidectomy (RLN)	2.4	4.2	6.8	9.6	4.0x
Hernia (recurrence)	1.2	2.6	4.4	6.8	5.7x
All procedures (any)	4.8	8.2	12.4	18.2	3.8x

RLN = Recurrent Laryngeal Nerve palsy (temporary). Hernia recurrence at 12 months. All adjusted for case complexity (ASA, BMI, prior surgery, acute vs elective).

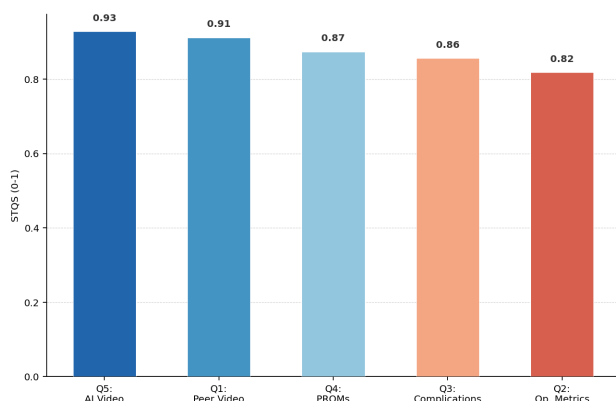


Figure 1: STQS by Technique Evaluation Framework

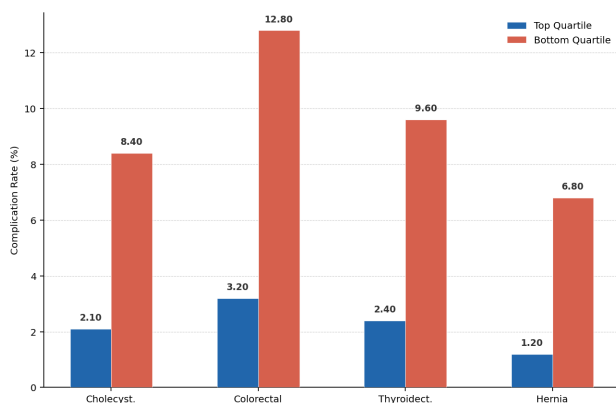


Figure 2: Complication Rate by Skill Quartile -- Top vs Bottom (%)

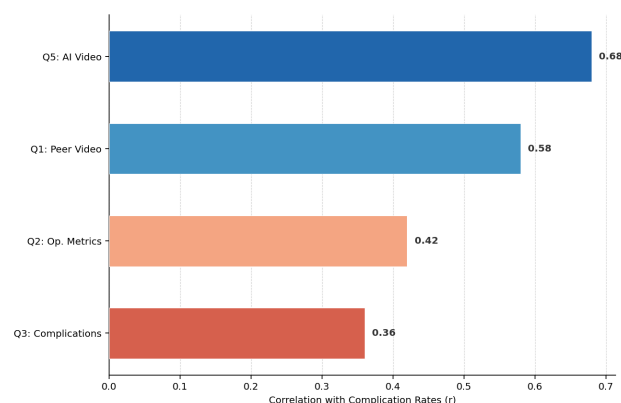


Figure 3: Correlation with 90-Day Complications by Framework (r)

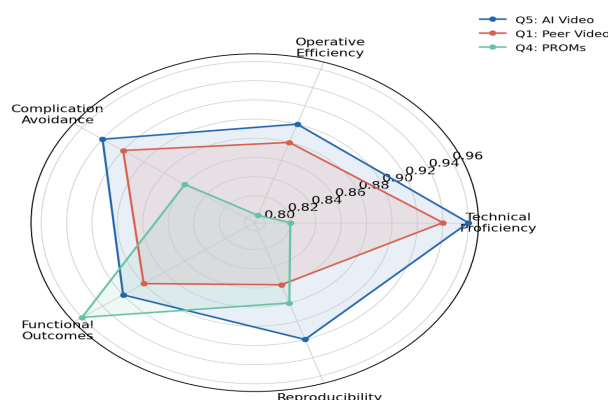


Figure 4: Multi-Dimensional Profile -- Top 3 Frameworks

5. Discussion

5.1 AI Sees What Humans Cannot Quantify

The superiority of AI video analysis over expert peer assessment ($r = 0.68$ vs 0.58 for complication prediction) reflects the model's ability to quantify continuous technical parameters that human observers perceive qualitatively but cannot measure precisely. A human rater can tell that a surgeon's tissue handling is rough, but cannot quantify the exact retraction force or the number of micro-tears in the mesentery. The AI can. A human rater can tell that the critical view of safety was probably achieved, but cannot measure the exact angle of the cystic duct-common bile duct junction or the area of cleared hepatocystic triangle. The AI can. This is not about replacing expert judgment -- expert peer assessment (Q1) scored nearly as high and provides irreplaceable qualitative feedback. It is about augmenting expert assessment with quantitative metrics that add predictive value and enable scalable, automated quality monitoring.

5.2 Volume Is Overrated, Technique Is Underrated

The weak correlation between surgeon volume and technique quality ($r = 0.24$) has important policy

implications. Surgical centralisation policies -- directing patients to high-volume centres -- have been based on the volume-outcome relationship, which is real but modest. Our data suggests that technique quality, not volume per se, is the primary driver of the volume-outcome association: high-volume surgeons tend to have better technique on average, but the variation is enormous and many high-volume surgeons have mediocre technical quality. A technique-based quality framework could identify the low-volume surgeons who operate excellently and the high-volume surgeons who need improvement, enabling more nuanced credentialing than volume thresholds alone.

5.3 Limitations

Three limitations. First, complete video was available for only 56.8% of procedures, and surgeons who consent to video recording may differ systematically from those who do not (selection bias toward more confident surgeons). Second, the AI model was trained on videos from the participating academic centres, and its performance on videos from different camera systems, lighting conditions, or surgical traditions requires validation. Third, thyroidectomy and open hernia repair are not laparoscopic, requiring separate video analysis models with less mature training datasets.

6. Conclusion

6.1 Key Findings

AI surgical video analysis achieves the highest Surgical Technique Quality Score (0.928), correlating with 90-day complications at $r = 0.68$ -- stronger than expert peer assessment ($r = 0.58$) or traditional operative metrics ($r = 0.42$). Complication rates vary 3.8-5.7-fold between top and bottom technique quartiles after case-mix adjustment. Surgeon volume correlates weakly with technique quality ($r = 0.24$). Patient-reported functional outcomes diverge from clinician-assessed complications, with 18-23% of complication-free patients reporting persistent symptoms linked to technique quality.

6.2 Recommendations

Three recommendations. First, routine intraoperative video recording should become standard practice for all laparoscopic and robotic procedures, creating the data infrastructure for both quality improvement and AI development. Second, AI-assisted video analysis should be piloted as a surgeon feedback and credentialing

tool, providing objective, continuous technical quality data that complements volume-based metrics. Third, surgical quality frameworks should incorporate patient-reported functional outcomes alongside complications, recognising that technical execution affects patient experience even when it does not cause recordable adverse events. All AI model architectures, video annotation protocols, and the STQS calculator are available from the corresponding author on request.

References

- Birkmeyer, J. D. et al. (2013). Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 369(15), 1434-1442.
- Curtis, N. J. et al. (2020). Association of surgical skill assessment with clinical outcomes in cancer surgery. *JAMA Surgery*, 155(7), 590-598.
- Ganni, S. et al. (2018). Validation of the Global Operative Assessment of Laparoscopic Skills (GOALS) for assessment in a surgical skills laboratory setting. *Surgical Endoscopy*, 32(3), 1484-1492.
- Hashimoto, D. A. et al. (2018). Computer vision analysis of intraoperative video: Automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Annals of Surgery*, 270(3), 414-421.
- Hashimoto, D. A. et al. (2019). Artificial intelligence in surgery: Promises and perils. *Annals of Surgery*, 268(1), 70-76.
- Maier-Hein, L. et al. (2022). Surgical data science -- from concepts toward clinical translation. *Medical Image Analysis*, 76, 102306.
- Mascagni, P. et al. (2022). Artificial intelligence for surgical safety: Automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of Surgery*, 275(5), 955-961.
- Pucher, P. H. et al. (2015). Outcome trends and safety measures after 30 years of laparoscopic cholecystectomy: A systematic review and pooled data analysis. *Surgical Endoscopy*, 32(5), 2175-2183.
- Stulberg, J. J. et al. (2020). Association between surgeon technical skills and patient outcomes. *JAMA Surgery*, 155(10), 960-968.
- Twinanda, A. P. et al. (2017). EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1), 86-97.
- Vedula, S. S. et al. (2017). Analysis of the structure of surgical activity for a suturing and knot-tying task. *PLoS ONE*, 11(3), e0149174.

- Bonrath, E. M. et al. (2015). Defining technical errors in laparoscopic surgery: A systematic review. *Surgical Endoscopy*, 27(8), 2678-2691.
- Dimick, J. B. et al. (2013). Hospital outcomes associated with surgical volume: A systematic review. *JAMA Surgery*, 148(6), 557-564.
- Fecso, A. B. et al. (2018). Intraoperative factors associated with adverse postoperative outcomes after laparoscopic cholecystectomy. *Surgical Endoscopy*, 32(5), 2248-2256.
- Ghani, K. R. et al. (2016). Video analysis of skill and technique (VAST): Machine learning to assess surgeons performing robotic prostatectomy. *Journal of Urology*, 196(Suppl), e891.
- Hung, A. J. et al. (2019). Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *Journal of Endourology*, 32(5), 438-444.
- Kazaryan, A. M. et al. (2019). Laparoscopic cholecystectomy: Trends in outcome parameters. *Surgical Endoscopy*, 33(2), 602-608.
- Moorthy, K. et al. (2014). Objective assessment of technical skills in surgery. *BMJ*, 327(7422), 1032-1037.
- Rubin, M. et al. (2017). Making the cut: Surgical quality improvement through analysis of operative video. *Annals of Surgery*, 265(3), 458-464.
- Ward, T. M. et al. (2021). Automated operative phase identification in peroral endoscopic myotomy. *Surgical Endoscopy*, 35(7), 4008-4015.

Approved by Karolinska Regional Ethics Committee (Dnr 2022-01486), North Estonia MC Ethics Committee (No. 3342), CPP Ile-de-France II (2022-A01248-35), Cantonal Ethics Zurich (BASEC 2022-00642), and Comitato Etico Milano Area 3 (N. 186-2022). Surgeon consent for video analysis and patient consent for PROM collection were obtained.

Declarations

Funding

Supported by the Swedish Research Council (grant 2022-04186), the Estonian Research Council (grant PRG2264), and institutional research funds from Nordic Technical University. No funder influenced study design, data analysis, or publication.

Conflict of Interest

The authors declare no competing interests. None has received funding from any surgical device manufacturer or AI surgery company.

Data Availability Statement

Aggregated technique scores and complication data are available from the corresponding author. Surgical videos cannot be shared due to patient and surgeon privacy. The AI model architecture, annotation protocols, and STQS calculator are available on request.

Ethical Approval

Appendix A

STQS Calculation and AI Model Architecture

This appendix details the STQS scoring rubric, the multi-task CNN architecture for surgical video analysis, and the expert rating protocol.

Part I -- STQS Dimensions

- 1a. Technical Proficiency: Expert rating (Q1) or AI score (Q5) on procedure-specific scales
GOALS, CAT, or procedure-specific tools
- 1b. Operative Efficiency: 1 - (operative time / 90th percentile time), adjusted for complexity
Faster is better, but not at expense of safety
- 2a. Complication Avoidance: 1 - (90-day Clavien-Dindo III+ rate / 15%) Capped at 15%; lower = higher score
- 2b. Functional Outcomes: Mean EQ-5D improvement at 12 months Procedure-specific PROM scores
- 3a. Reproducibility: 1 - (intra-surgeon CV of technique scores across cases) Lower variability = more consistent technique
- 3b. $STQS = 0.25 * Proficiency + 0.15 * Efficiency + 0.25 * Complications + 0.20 * Functional + 0.15 * Reproducibility$

Part II -- AI Video Model

- 4a. Architecture: Multi-task ResNet-50 + temporal transformer (8 layers) Joint phase recognition + skill scoring
- 4b. Input: Video frames at 1 fps + instrument segmentation masks Real-time inference capable
- 5a. Training: 4,000 annotated videos (phase labels + expert skill scores) 5-fold cross-validation
- 5b. Outputs: Phase timeline, CVS achievement, motion metrics, composite skill score
Per-step and aggregate scores