Surgical Endoscopy (2022) 36:7938–7948
https://doi.org/10.1007/s00464-022-09296-6

REVIEW ARTICLE

# The association between video-based assessment of intraoperative technical performance and patient outcomes: a systematic review

**Saba Balvardi[1,2] · Anitha Kammili[1,2] · Melissa Hanson[1,2] · Carmen Mueller[1,2] · Melina Vassiliou[1,2] · Lawrence Lee[1,2] · Kevin Schwartzman[3,4] · Julio F. Fiore Jr.[1,2] · Liane S. Feldman[1,2]**

Received: 18 October 2021 / Accepted: 18 April 2022 / Published online: 12 May 2022


## Abstract

**Background** Efforts to improve surgical safety and outcomes have traditionally placed little emphasis on intraoperative performance, partly due to difficulties in measurement. Video-based assessment (VBA) provides an opportunity for blinded and unbiased appraisal of surgeon performance. Therefore, we aimed to systematically review the existing literature on the association between intraoperative technical performance, measured using VBA, and patient outcomes.

**Methods** Major databases (Medline, Embase, Cochrane Database, and Web of Science) were systematically searched for studies assessing the association of intraoperative technical performance measured by tools supported by validity evidence with short-term (≤30 days) and/or long-term postoperative outcomes. Study quality was assessed using the Newcastle–Ottawa Scale. Results were appraised descriptively as study heterogeneity precluded meta-analysis.

**Results** A total of 11 observational studies were identified involving 8 different procedures in foregut/bariatric ($n=4$), colorectal ($n=4$), urologic ($n=2$), and hepatobiliary surgery ($n=1$). The number of surgeons assessed ranged from 1 to 34; patient sample size ranged from 47 to 10,242. High risk of bias was present in 5 of 8 studies assessing short-term outcomes and 2 of 6 studies assessing long-term outcomes. Short-term outcomes were reported in 8 studies (i.e., morbidity, mortality, and readmission), while 6 reported long-term outcomes (i.e., cancer outcomes, weight loss, and urinary continence). Better intraoperative performance was associated with fewer postoperative complications (6 of 7 studies), reoperations (3 of 4 studies), and readmissions (1 of 4 studies). Long-term outcomes were less commonly investigated, with mixed results.

**Conclusion** Current evidence supports an association between superior intraoperative technical performance measured using surgical videos and improved short-term postoperative outcomes. Intraoperative performance analysis using video-based assessment represents a promising approach to surgical quality-improvement.

**Keywords** Video-based assessment · VBA · Intraoperative performance · Intraoperative assessment tools · Surgical outcome

Evidence supports that 40–60% of adverse events in surgical patients can be linked to errors in the operating room [1–3]. Yet efforts to improve surgical outcomes have largely focused on perioperative care with very little emphasis on measuring and improving operative performance [4]. Difficulty in accessing information on 'what happens in the operating room' and lack of appropriate tools for assessment of intraoperative performance have hampered this area of research [4, 5]. However, the expansion of image-guided surgery including laparoscopic and robotic operations facilitates capture, storage, and sharing of recorded procedures. Consequently, video-based assessment (VBA) may provide a valuable opportunity to measure intraoperative performance

✉ Liane S. Feldman
liane.feldman@mcgill.ca

1 Department of Surgery, McGill University, 1650 Cedar Ave, D6-136, Montreal, QC H3G 1A4, Canada

2 Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, QC, Canada

3 Respiratory Division, Department of Medicine, McGill University, Montreal, QC, Canada

4 McGill International Tuberculosis Centre, Research Institute of the McGill University Health Centre, Montreal, QC, Canada

while minimizing observer bias related to unblinded in-theater evaluations [6, 7].

There is significant interest in the use of VBA of intraoperative performance for formative assessment in education and coaching [4, 8, 9]. In addition, there is interest in the use of VBA for summative 'high stakes' decisions such as certification after completion of surgical training [5] or after learning a new procedure [10, 11]. However, the use of VBA to inform competency decisions for trainees requires robust supporting evidence. A landmark paper from Birkmeyer et al. published in 2013 reported a significant association between surgeon technical performance and outcomes after Roux-en-Y gastric bypass, including complications, reoperations, and readmissions [12]. A systematic review, however, identified important limitations in the literature published in this field related to lack of standardized assessment tools and reliance on indirect observations of technical performance such as postoperative imaging or pathological specimen quality [13]. This has become an active area of research and several studies published subsequent to that review contributed new evidence that may further inform the integration of VBA into credentialing, certification, coaching, and quality improvement processes for practicing surgeons. Therefore, the objective of this study was to systematically review and summarize the existing literature on the association between intraoperative technical performance measured using VBAs and patient outcomes.

## Materials and methods

This review was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) [14]. The review protocol was registered a priori at Open Science Framework (osf.io/c29yb).

### Eligibility criteria

We included studies that (1) measured intraoperative technical performance of practicing surgeons from recorded cases; (2) described the association of intraoperative technical performance with the outcomes of patients undergoing the same type of procedure; and (3) used a performance assessment tool with published validity evidence supporting their intended use and interpretation. Studies from all surgical specialities published after 1990 (introduction of image-guided procedures) [15] were included. Exclusion criteria included (1) studies evaluating surgical trainees; (2) studies that relied solely on surrogate measures of technical performance such as postoperative imaging or pathological specimen; (3) studies with qualitative assessment of intraoperative technical performance only (i.e., lack of a standardized assessment tool); (4) case reports, comments, editorials, and non-human studies; and (5) abstracts that could not be traced to full-text articles. There were no language restrictions.

### Literature search

The following databases were searched for relevant studies: Medline (via OvidSP and PubMed [for articles ahead of print]), Embase (OvidSP), The Cochrane Database (via Cochrane Library, including Cochrane Central Register of Controlled Trials, Database of Abstracts of Reviews of Effects, and National Health Service Economic Evaluation Database), and Web of Science (Thomson Reuters). The search strategies (**eMethods 1**) were developed by an experienced medical librarian according to the best practice recommendations [16]. The reference list of the selected studies was screened for further studies that met the inclusion criteria. [17] Searches were carried out in August 2020 and updated in March 2021 before manuscript submission. No language restrictions were applied.

### Study selection and data extraction

Two reviewers (SB and AK) independently assessed titles, abstracts, and selected full texts of the articles obtained through the literature review. Any discrepancies between the included and excluded articles were resolved by consensus between the reviewers or by consulting a third independent reviewer (MH).

### Quality assessment of individual studies

The methodological quality for each study included in the final selection was independently judged by two reviewers (SB and AK) using the Newcastle–Ottawa Scale (NOS) [18]. Any discrepancies were resolved by consensus between the reviewers or by consulting a third independent reviewer (LF). NOS is a validated system developed for the assessment of quality of non-randomized trials based on three domains: selection of the study groups (maximum of 4 stars), comparability of the groups (maximum of 2 stars), and ascertainment of the exposure or outcome of interest (maximum of 3 stars) with a maximum total score of 9 stars [19]. Although there are no defined cutoff values differentiating high-quality from low-quality study methods in the NOS tool, studies with fewer than 6 stars or with 1 star for the selection of participants or outcome ascertainment, or zero for any domain were deemed to have high risk of bias. [20–23] We followed a priori criteria for risk of bias analysis based on the NOS guidelines, as outlined in Supplemental Digital Content 1. [24, 25].

## Data synthesis

This systematic review was reported using a narrative synthesis approach [26]. Meta-analysis was precluded as the identified studies were heterogeneous with respect to population, exposure, and outcome measures.

## Results

A total of 3984 unique articles were identified and 31 articles were chosen for final full-text review after screening of titles and abstracts (Fig. 1). There were 3 additional studies identified through other sources (cross referencing [$n=2$] [27, 28] or expert suggestions of recent papers which had not yet been indexed in Medline [$n=1$] [29]. Twenty-three articles were excluded (articles and reasons for exclusion are listed in Supplemental Digital Content 1) and 11 articles met eligibility criteria. [12, 29–38].

Characteristics of the included studies are summarized in Table 1. All were observational studies (10 cohort and 1 case–control study). All the other ten identified studies followed after the publication of the landmark paper by Birkmeyer et al. [12] Eight of 11 studies were multicenter collaborations. Two studies involved urologic procedures [34, 36] with the remainder involving general surgery procedures (foregut/bariatric [$n=4$], colorectal [$n=4$] and hepatobiliary surgery [$n=1$]) [12, 29–33, 35, 37, 38]. Eight different procedures were evaluated in these studies. All studies involved minimally invasive surgical procedures (two studies in robotic surgery and 9 in laparoscopic surgery). The number of surgeons evaluated in each study ranged from 1 to 34. The rate of participation of invited surgeons ranged from 32 to 100% when specified. A range of 47–10,242 patients were assessed for surgical outcomes in the identified studies.

Table 2 summarizes the characteristics of the intraoperative technical performance assessment tools used and the features of the study designs that may influence their uses and interpretations [13]. A wide variety of generic and procedure-specific assessment tools were used, with 54% of the studies ($n=6$) using the generic modified Objective Structured Assessment of Technical Skills (mOSATS) tool. The Generic Error Rating Tool (GERT) was the only error rating tool identified which was used in two identified studies. The remaining assessment tools used in these studies were procedure-specific, including the American Society of Colon and Rectal Surgeons (ASCRS) Video Assessment Tool, which was used in two out of three of the studies evaluating laparoscopic colectomy. Six studies assessed only critical parts of a given procedure that were defined a priori that included parts of an operation such as the anastomosis or critical dissections. Three out of these six studies, the intraoperative recording was edited by the research team to only include the a priori identified critical section of the operation. Five studies involved VBA of the entire procedure. In ten studies, the assessors were blinded to the patient and surgeon identifiers, and in one study, this was not specifically reported. Eight studies characterized the assessors as "expert," while

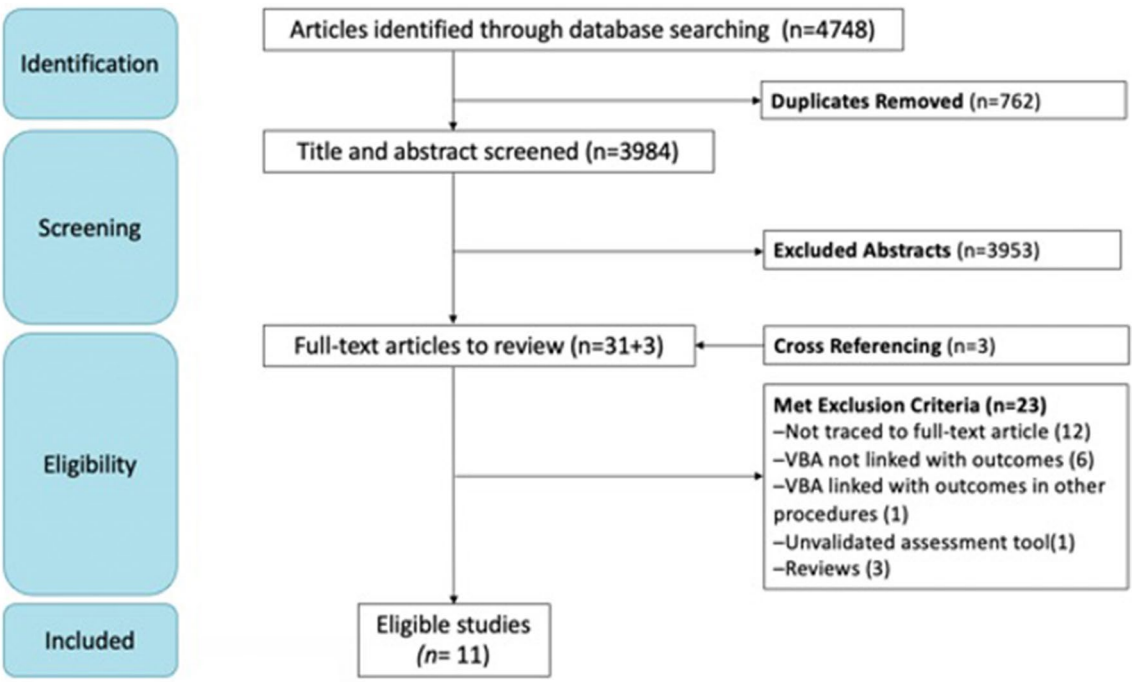


**Fig. 1** PRISMA flow diagram. (PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses) [14]

**Table 1** Overview of the included studies

| Author | Year | Design | Country | Surgeons *n* (%) [a] | Patients *n* | Specialty | Operation Assessed | Primary Outcomes | Secondary outcomes |
|---|---|---|---|---|---|---|---|---|---|
| Varaban et al. [30] | 2021 | Multicenter Retrospective Cohort | United States | 25 (35%) | 3502 | GS | Lap sleeve gastrectomy | Complications (30 d) | Readmission (30 d)<br>Reoperation (30 d)<br>ED visits (30 d)<br>EBWL % (1 year) |
| Brajcich et al. [29] | 2020 | Multicenter retrospective cohort | United states | 15 (NS) | 609 | GS | Lap right hemicolectomy | Survival (5-year) | Nil |
| Stulberg et al. [31] | 2020 | Multicenter prospective cohort | United States | 17 (NS) | 1120 | GS | Lap Right Hemicolectomy | Complications (30 d)<br>Mortality (30 d)<br>Readmission (30 d)<br>Reoperation (30 d) | Skill-related morbidity (30 d)[b]<br>Skill-unrelated morbidity (30 d)[b] |
| Curtis et al. [32] | 2020 | Multicenter prospective cohort | Australia, New Zealand and United Kingdom | 34 (100%) | 176 | GS | Lap total Mesorectal excision | Complications (30 d)<br>Reoperation (30 d)<br>Readmission (30 d) | Overall Survival (2–4 years)<br>Cancer Recurrence (2–4 years) |
| Fecso et al. [33] | 2019 | Multicenter retrospective cohort | Canada | 3 (10%) | 61 | GS | Lap gastrectomy | Complications (30 d) | Nil |
| Goldenberg et al. [34] | 2017 | Single center prospective case–control | Canada | 1 (100%) | 47 | Urology | Robotic assisted radical prostatectomy | Continence (3 mo) | Nil |
| Scally et al. [35] | 2016 | Multicenter retrospective cohort | United States | 20 (27%) | 3631 | GS | Lap gastric bypass | EBWL % (1 year) | Patient satisfaction (1 year) |
| Paterson et al. [36] | 2016 | Single center prospective cohort | Scotland | 1 (100%) | 200 | Urology | Extraperitoneal lap prostatectomy | Continence (3 mo) | Continence (12 mo)<br>Readmission (30, 90 and 120 d)[c]<br>Reoperation (30, 90 and 120 d)[c]<br>ED visits (30, 90 and 120 d)[c]<br>Complications (30, 90 and 120 d)[c]<br>Erectile dysfunction (90 d)[c] |
| Hogg et al. [37] | 2016 | Single center retrospective cohort | United States | NS | 133 | GS | Robotic whipple | Postoperative pancreatic fistula | Nil |
| MacKenzie et al. [38] | 2015 | Multicenter prospective cohort | United Kingdom | 20 (32%) | 171 | GS | Lapright and left hemicolectomy | Surgical complications (30 d) | Nil |
| Birkmeyer et al. [12] | 2013 | Multicenter retrospective cohort | United States | 15 (NS) | 10,242 | GS | Lap gastric bypass | Complications (30 d) | Mortality (30 d)<br>readmission (30 d)<br>reoperation (30 d)<br>ED visits (30 d) |

*NS* not specified, *GS* general surgery, *Lap* laparoscopic; *ED* emergency department, *EBWL%* excess body weight loss %

[a]Number of surgeons assessed (*n*) and proportion of surgeons asked to participate who agreed to participate (%)

[b]These composite outcome groups were created a priori by the authors to reflect outcomes that conceptually should or should not be related to a surgeon's technical skill

[c]No effect size was reported for these a priori outcomes, and therefore, they were excluded from the following analysis

three studies characterized them as "peer assessors." Only six (54%) studies described any attempt to train or calibrate the raters in using the assessment rubrics. Videos used for intraoperative technical performance assessment were submitted in two methods. In five studies, participating surgeons chose and submitted one video as representative of their

**Table 2** Overview of intraoperative skills assessment

| Author | Assessment tool | Part of operation assessed | Rater qualification | Blinded assessment | Edited video | Rater training[a] | Video submission |
|---|---|---|---|---|---|---|---|
| Varaban et al. [30] | Modified OSATS [b] | Whole procedure | Peer raters | Yes | No | No | 1–2 videos submitted by each surgeon |
| Brajcich et al. [29] | ASCRS[c] video assessment tool | NA | Peer raters & experts | NA | NA | NA | 1 video submitted by each surgeon |
| Stulberg et al. [31] | Combination of OSATS and ASCRS video assessment tool | Whole procedure | Peer raters & experts | Yes | NA | Yes | 1 video submitted by each surgeon |
| Curtis et al. [32] | LapTMEpt performance assessment tool | Whole procedure | Expert | Yes | NA | Yes | 1 video per patient |
| Fecso et al. [33] | OSATS<br>GERT[d] | Critical parts of procedure | Experts | Yes | No | Yes | 1 video per patient |
| Goldenberg et al. [34] | GEARS[e]<br>GERT | Whole procedure | Experts | Yes | No | Yes | 1 video per patient |
| Scally et al. [35] | Modified OSATS | Critical parts of procedure | Peer raters | Yes | Yes | No | 1 video submitted by each surgeon |
| Paterson et al. [36] | VELP-score[f] | Critical part of procedure | Experts | Yes | Yes | NA | 1 video per patient |
| Hogg et al. [37] | Modified OSATS<br>Technical scoring pancreaticojejunostomy | Critical parts of procedure | Experts | Yes | No | Yes | 1 video per patient |
| MacKenzie et al. [38] | Competency assessment tool | Whole procedure | Experts | Yes | No | Yes | 1 video per patient |
| Birkmeyer et al. [12] | Modified OSATS | Critical parts of procedure | Peer raters | Yes | Yes | No | 1 video submitted by each surgeon |

*NA* not available

[a]Any attempt at training

[b]Objective structured assessment of technical skills

[c]American Society of Colon and Rectal Surgeons

[d]Generic error rating tool

[e]Global evaluative assessment of robotic skill

[f]Video recorded extraperitoneal laparoscopic radical prostatectomy score

overall performance. In this approach, the surgeon's technical performance was estimated from that single video, and patient outcomes for each surgeon were determined from an existing registry. In the remaining six studies, videos were available for each case, and the association between intraoperative technical performance and outcomes was analyzed for each patient.

Quality assessment of each study was performed using the NOS tool [19]. A total of 6 studies were deemed to have low risk of bias and 5 studies to have high risk of bias (Table 3). A common reason for penalizing the quality of the studies was bias in selection of participants in the study ($n=8$) [12, 29–31, 33, 35, 37, 38] followed by bias in measurement of exposure and non-disclosure of frequency and handling of missing data ($n=10$) [12, 29–33, 35–38]. A complete description of the risk of bias assessment for each study is reported in Supplemental Digital Content 1.

The relationship between intraoperative technical performance and postoperative outcomes for each study is summarized in Table 4. The outcomes assessed were categorized as short-term ($\leq 30$ days) or long-term ($>30$ days). Short-term outcomes (including 30-day complications, reoperations, readmissions, emergency department visits, and survival) were reported in 8 studies. Better intraoperative performance was associated with fewer postoperative complications (6 of 7 studies) in laparoscopic right and left hemicolectomy, laparoscopic total mesorectal excision, laparoscopic gastrectomy, laparoscopic gastric bypass, and robotic Whipple procedures. Out of these seven studies, three had low risk of bias [12, 30, 31], 2 of which demonstrated an association

**Table 3** Study quality assessment for primary outcomes

| Author | Design | The Newcastle–Ottawa scale[a] | | | Risk of bias[b] |
|---|---|---|---|---|---|
| | | Selection | Comparability | Outcome/exposure | |
| Varaban et al. [30] | Multicenter retrospective cohort | ☆☆ | ☆☆ | ☆☆ | Low Risk of Bias |
| Brajcich et al. [29] | Multicenter retrospective cohort | ☆☆ | ☆☆ | ☆☆ | Low Risk of Bias |
| Stulberg et al. [31] | Multicenter prospective cohort | ☆☆ | ☆☆ | ☆☆ | Low Risk of Bias |
| Curtis et al. [32] | Multicenter prospective cohort | ☆☆☆ | | ☆☆ | High Risk of Bias |
| Fecso et al. [33] | Multicenter retrospective cohort | ☆☆ | ☆☆ | ☆ | High Risk of Bias |
| Goldenberg et al. [34] | Single center prospective case–control | ☆☆☆ | ☆☆ | ☆☆☆ | Low Risk of Bias |
| Scally et al. [35] | Multicenter retrospective cohort | ☆☆☆ | ☆☆ | ☆☆ | Low Risk of Bias |
| Paterson et al. [36] | Single center prospective cohort | ☆☆☆ | ☆☆ | ☆ | High Risk of Bias |
| Hogg et al. [37] | Single center retrospective cohort | ☆☆☆ | ☆☆ | ☆ | High Risk of Bias |
| MacKenzie et al. [38] | Multicenter prospective cohort | ☆☆☆☆ | ☆☆ | ☆ | High Risk of Bias |
| Birkmeyer et al. [12] | Multicenter retrospective cohort | ☆☆ | ☆☆ | ☆☆ | Low Risk of Bias |

[a]Maximum number of starts are 9 (4 Selection, 2 comparability, 3 outcome/exposure)

[b]Studies with less than 6 stars or with one star for the selection of participants or outcome ascertainment, or zero for any domain were deemed to have high risk of bias

between better intraoperative performance and fewer postoperative complications (rate reduction between 9.2% and 5.1%) [12, 31]. Better intraoperative performance was associated with fewer reoperations in 3 of 4 studies (rate reduction between 0.7% and 2.5%), including all 3 studies with low risk of bias [12, 30–32].Better intraoperative performance had an association with fewer readmission in only 1 of 4 studies [12]; only one of these studies (that showed no association) had a high risk of bias [31]. All studies looking at ED visits and mortality were of low risk of bias [12, 30, 31]. One of 2 studies showed an association between better intraoperative performance and lower ED visits and mortality. [12].

The impact of intraoperative performance on long-term outcomes was reported in 6 studies and supported by studies focused on weight loss (1 of 2 studies, both with low risk of bias) [30, 35], and patient satisfaction (1 of 1 study with high risk of bias) [35], but not cancer recurrence (0 of 1 study with high risk of bias) [32]. Cancer survival was investigated in 2 studies: an association between better intraoperative technical performance and longer overall cancer survival was supported by one study with low risk of bias [29] with a second study with high risk of bias reporting a large but non-statistically significance increase in overall survival. [32] In minimally invasive prostatectomy, an association between intraoperative technical performance and improved 3 month postoperative urinary continence rate was supported in 2 studies (1 with low risk of bias [34] and one with high risk of bias [36]) (Table 4). Four studies reported the association between intraoperative technical performance and pathological outcomes. [29, 32, 36, 38] Of the 3 studies investigating the association between intraoperative technical performance and lymph node yield, 2 showed no association [29, 32] and 1 showed a significant association (13 vs. 18 LNs in colon cancer) [38]. One study showed a significant association between better intraoperative technical performance and higher rate of pathologic success in rectal cancer surgery (defined as mesorectal fascial plane, circumferential margin $\geq$ 1 mm, and distal margin $\geq$ 1 mm) [32] and another reported an association with the distal margin in left colon cancer surgery (median 3 vs. 4 cm). [38].

## Discussion

This systematic review summarizes the existing literature investigating the association between intraoperative technical performance, as evaluated using VBA measures, and patient outcomes. Despite study heterogeneity, the results support the association between better intraoperative technical performance and improved short-term outcomes including 30-day complications and reoperations in laparoscopic colectomy, laparoscopic total mesorectal excision, laparoscopic gastrectomy, laparoscopic gastric bypass, and robotic Whipple procedures. There was more limited evidence supporting the relationship between technical performance and short-term resource utilization (readmissions and ED visits), as well as longer-term outcomes such as weight loss after bariatric surgery and survival after cancer resections.

Our study builds on the previous systematic review assessing the association between technical performance and patient outcome, which included studies conducted up to 2014. The earlier review, which included 24 studies, included only one study where an intraoperative assessment

**Table 4** Association of intraoperative performance with postoperative outcomes

| Author | Quality assessment | Operation assessed | Outcomes assessed (duration) | Effect |
|---|---|---|---|---|
| Varaban et al. [30] | Low risk of bias | Lap Sleeve Gastrectomy | Complications (30 d)<br>Readmission (30 d)<br>Reoperation (30 d)<br>ED visits (30 d)<br>EBWL % (1 year) | Rho: 0.21, $p=0.30$<br>Rates: 1.9% vs. 2.9%, $p=0.25$<br>Rates: 0.2% vs. 0.9%, $p<0.0001$*<br>Rates: 8.6% vs. 8.2%, $p=0.57$<br>58.8% vs. 56.1%, $p<0.03$* |
| Brajcich et al. [29] | Low risk of bias | Lap right hemicolectomy | Survival (5-year) | HR: 0.31 [0.18, 0.54]* |
| Stulberg et al. [31] | Low risk of bias | Lap hemicolectomy | Complications (30 d)<br>Mortality (30 d)<br>Readmission (30 d)<br>Reoperation (30 d)<br>Skill-related morbidity (30 d)<br>Skill-unrelated morbidity (30 d) | MD: 5.1% [0.4%, 9.8%], $p=0.03$*<br>MD: 0.3% [–0.4%, 0.9%], $p=0.59$<br>MD: 1.5% [−0.9%, 4.0%], $p=0.27$<br>MD: 2.5% [0.5%, 4.6%], $p=0.02$*<br>MD: 6.5% [−0.8%, 13.8%], $p=0.08$<br>MD: 2.9% [−4.2%, 9.9%], $p=0.55$ |
| Curtis et al. [32] | High risk of bias | Lap Total Mesorectal Excision | Complications (30 d)<br>Reoperation (30 d)<br>Readmission (30 d)<br>Overall Survival (2–4 years)<br>Cancer Recurrence (2–4 years) | Rates: 23.3% vs. 55.3%, $p=0.03$*<br>Rates: 3.3% vs. 6.3%, $p=0.6$<br>Rates: 0% vs. 33.3%, $p=0.19$<br>Rates: 8.7% vs. 96.6%, $p=0.46$<br>Rates: 70.0% vs. 74.2, $p=0.46$ |
| Fecso et al. [33] | High risk of bias | Lap gastrectomy | Complications (30 d) | Rho: 0.401, $p=0.001$* |
| Goldenberg et al. [34] | Low risk of bias | Robotic assisted radical prostatectomy | Continence (3 month) | OR = 0.55 [0.33, 0.91]* |
| Scally et al. [35] | Low risk of bias | Lap gastric bypass | EBWL % (1 year)<br>Patient satisfaction (1 year) | 67.2% vs. 68.5%; $p=0.86$<br>IR: 90.3% vs. 87.1%; $p=0.05$* |
| Paterson et al. [36] | High risk of bias | Extraperitoneal lap prostatectomy | Continence (3 month)<br>Continence (12 months) | HR: 7.3 [2.2, 24.6] vs 10.9 [2.0,39.5] vs. 5.5 [1.5, 19.9]* for decreasing skill level<br>HR: 5.0 [1.2, 22.0] vs. 10.9 [2.01, 40.0] vs. 5.5 [1.4, 18.0] for decreasing skill level |
| Hogg et al. [37] | High risk of bias | Robotic whipple | Postoperative pancreatic fistula | OR: 0.82 [0.70, 0.96]* |
| MacKenzie et al. [38] | High risk of bias | Lap right and left hemicolectomy | Surgical complications (30 d) | RRR:0.68 [0.31, 0.85], $p=0.005$* |
| Birkmeyer et al. [12] | Low risk of bias | Lap gastric bypass | Complications (30 d)<br>Mortality (30 d)<br>Readmission (30 d)<br>Reoperation (30 d)<br>ED visits (30 d) | Rates: 5.2% vs. 14.5%, $p<0.001$*<br>Rates: 0.05% vs. 0.26%, $p=0.01$*<br>Rates: 2.7% vs. 6.3%, $p<0.001$*<br>Rates: 1.6% vs. 3.4%, $p=0.01$*<br>Rates: 3.8 vs. 10.2%, $p=0.004$* |

Data are presented according to the primary analysis reported in each study as MD (95% CI): Mean Difference, HR (95% CI): Hazard Ratio, RRR (95% CI): Relative Risk Reduction, Rates: Superior Skill group vs. Inferior Skill Group, Rho (p-value): Spearman Correlation Coefficient, OR (95% CI): Odds Ratio

*ED* emergency department, *EBWL%* excess body weight loss %

*Indicated statistical significance

tool with validity evidence was used for VBA of practicing surgeons, while the remaining studies relied on indirect evaluations of intraoperative performance such as postoperative imaging or pathological specimens [12, 13] Our systematic review was further strengthened with compliance with PRISMA methodological standards and the use of cross referencing to maximize our literature search. [14, 17].

Given that the majority of the VBA tools used in the studies, such as mOSATS, focus mostly on elements of psychomotor proficiency, such as dexterity and tissue handling, it is not surprising that associations were found between intraoperative performance and short-term safety outcomes, while associations with long-term efficacy outcomes were less clear. While intraoperative technical performance seems important in preventing early complications like bleeding and infection, most assessment tools used in the included studies do not fully capture the complex cognitive skills related to surgical expertise that may have a larger role to play in determining the long-term effectiveness of the operation [5, 39]. Therefore, the tool used for VBA should be selected based on the outcome of interest. An additional source of variability is that operations are not standardized between surgeons and these variations (e.g., oversewing versus not oversewing of the staple-line or length of the roux-limb in bariatric surgery) may also be associated with postoperative outcomes [40, 41]. However, technical variation was not considered in any of the identified studies in this review, which may also contribute to the heterogeneity observed in the effect measures [30]. One of the long-term outcomes that was associated with superior intraoperative technical performance was improved cancer survival in 2 studies, despite the mixed findings in the association between intraoperative performance and pathology outcomes. This may be related to the detrimental impact of major early postoperative complications on oncological outcomes related to increased systematic spread or delayed adjuvant treatment. [42–44].

The association between surgeon technical performance and patient outcome has several important implications. It suggests a potential avenue for quality improvement and continuing professional development through feedback, benchmarking, and coaching [8, 45]. Similarly, there is an interest in using VBA to measure and improve surgical techniques from leading groups such as the American Board of Surgery [46]. It is important to highlight that association does not imply causation; while there is evidence for the benefits of video analysis and feedback in surgical trainees [47], additional studies are required to support the effectiveness of this approach for practicing surgeons. Additionally, for VBA to be used to inform higher-stakes decisions (e.g., certification and credentialing), the measurement tools need to be supported by rigorous studies supporting their validity for that use and be representative of all domains the tool seeks to measure including operative safety and effectiveness [5, 48]. There is limited evidence supporting the use of the generic assessment tools identified in this review for summative video-based evaluation in practicing surgeons [49, 50]. However, other instruments identified in our study were in fact developed specifically to assess performance of a specific procedure by practicing surgeons, using a recorded case, with evidence provided supporting their uses, interpretations, and psychometric properties. [10, 32] This work is critical as automated metrics of performance using computer vision and machine learning are rapidly being developed [51]. Finally, the ability to accurately document and measure variations in surgical technique using VBA has implications for surgical research, with many randomized trials now requiring submission and analysis of procedure video to ensure quality and standardization. [52].

We identified significant heterogeneity in study design related to video editing, the type of assessment tool, rater qualification, and rater training. These characteristics were selected based on the published recommendations for minimizing measurement error when using VBAs. [13, 53] Although our review only included studies using assessment tools supported by validity evidence, evaluation of the strength of the validity evidence for the intended uses and interpretations falls outside the scope of this review. As discussed earlier, the development and use of assessment tools with robust psychometric properties should be standard practice for video-based evaluations. [48].

While most studies followed the recommendation to use blinded evaluators, rater qualification varied and was either described as "peer" or "expert" evaluation. The definition of expert raters varied between studies but was commonly described as an experienced surgeon in the field (i.e., some may argue this is a "peer") or as having familiarity in the use or development of intraoperative assessment tools (i.e., may not be clinical expertise). Use of multiple peer raters (as opposed to defined experts in the field) has been justified in the literature based on the theory that the collective intelligence of a group may solve problems more efficiently than individuals [7]. The literature supporting peer VBA assessment in comparison to expert assessment (the default gold standard) has been mixed [54, 55] with supporting evidence for their use in evaluating simple tasks such as knot tying [56] and in the presence of added information such as intraoperative audio [55]. Since the use of peer assessors would significantly increase the feasibility of larger scale assessment programs, the qualifications of the raters should be better defined, and evidence to support optimizing rater training should be prioritized in future studies.

There was also wide range of definitions for rater training, ranging from passive training based on descriptive manuals [32] to full training programs with continuous calibration of the assessors [33]. Only one of 5 studies that used peer assessors described any attempts at rater training. In studies with lack of peer training, lack of familiarity with the nuances of assessment tools can result in non-differential measurement error, resulting in underestimation of the effect size and biasing the analysis toward the null. For future studies, rater training is recommended to enhance reliability and reduce non-differential measurement bias, but more work

is needed to determine the optimal mode of rater training. [7, 13, 57].

Inconsistency in the association between intraoperative technical performance and outcomes between studies may be related to other issues in study design. In almost half of the studies, a single video chosen by the participating surgeon was used, compared to the alternative of having one video per patient. The former method is not only susceptible to selection bias, but also evaluating a surgeon based on a single video does not take into account a surgeon's learning curve or the evolution of their technique throughout their years of practice. On the other hand, surgeons would likely select their "best" videos which would bias the results toward the null. The number of assessments required for a reliable score using VBA has been investigated in trainees; however, this information is lacking in assessment of practicing surgeons. [58].

This review has several limitations. Study heterogeneity precluded meta-analysis. In addition to the risk of measurement bias discussed above, eight of the eleven identified studies were at high risk of selection bias. The most common reason was the degree of participation from surgeons, consistently reported below 35% of invited participants. Another area of potential bias was the inclusion of patients based on the availability of intraoperative videos versus having a consecutive cohort of patients where video and outcome data were both available for every patient. Twelve abstracts were excluded because they were not yet traced back to a full-text article. Our systematic review also did not identify any studies of open surgical procedures likely due to increased complexity for recording.

This review contributes evidence regarding the relationship between technical performance as measured through video-based assessment and surgical outcomes, supporting the association between superior intraoperative technical performance and lower risk of perioperative complications and reoperations. Long-term outcomes were less commonly investigated, with mixed results. Future research should investigate the impact of technical performance and technical variation on postoperative outcomes in a more diverse range of procedures and investigate the effectiveness of interventions to improve technical skill on patient outcomes.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00464-022-09296-6.

**Funding** Fonds de la recherche en Sante du Quebec (FRSQ)- Grant Number 288097.

## Declarations

**Disclosures** Saba Balvardi, Anitha Kammili, Melissa Hanson, Carmen Mueller, Melina Vassiliou, Lawrence Lee, Kevin Schwartzman, Julio F Fiore Jr., Liane S Feldman declare that they have no conflict of interest.

## References

1. Zegers M, de Bruijne MC, Wagner C et al (2009) Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care 18(4):297–302. https://doi.org/10.1136/qshc.2007.025924
2. Kable AK, Gibberd RW, Spigelman AD (2002) Adverse events in surgical patients in Australia. Int J Qual Health Care 14(4):269–276. https://doi.org/10.1093/intqhc/14.4.269
3. Fabri PJ, Zayas-Castro JL (2008) Human error, not communication and systems, underlies surgical complications. Surgery 144(4):557–63; discussion 563–5. https://doi.org/10.1016/j.surg.2008.06.011
4. Dimick JB, Varban OA (2015) Surgical video analysis: an emerging tool for improving surgeon performance. BMJ Qual Saf 24(8):490–491. https://doi.org/10.1136/bmjqs-2015-004439
5. Feldman LS, Pryor AD, Gardner AK et al (2020) SAGES video-based assessment (VBA) program: a vision for life-long learning for surgeons. Surg Endosc 34(8):3285–3288. https://doi.org/10.1007/s00464-020-07628-y
6. Yanes AF, McElroy LM, Abecassis ZA, Holl J, Woods D, Ladner DP (2016) Observation for assessment of clinician performance: a narrative review. BMJ Qual Saf 25(1):46–55. https://doi.org/10.1136/bmjqs-2015-004171
7. Bilgic E, Valanci-Aroesty S, Fried GM (2020) Video assessment of surgeons and surgery. Adv Surg 54:205–214. https://doi.org/10.1016/j.yasu.2020.03.002
8. Greenberg CC, Dombrowski J, Dimick JB (2016) Video-based surgical coaching: an emerging approach to performance improvement. JAMA Surg 151(3):282–283. https://doi.org/10.1001/jamasurg.2015.4442
9. Grenda TR, Pradarelli JC, Dimick JB (2016) Using surgical video to improve technique and skill. Ann Surg 264(1):32–33. https://doi.org/10.1097/SLA.0000000000001592
10. Mackenzie H, Cuming T, Miskovic D et al (2015) Design, delivery, and validation of a trainer curriculum for the national laparoscopic colorectal training program in England. Ann Surg 261(1):149–156. https://doi.org/10.1097/SLA.0000000000000437
11. Mori T, Kimura T, Kitajima M (2010) Skill accreditation system for laparoscopic gastroenterologic surgeons in Japan. Minim Invasive Ther Allied Technol 19(1):18–23. https://doi.org/10.3109/13645700903492969
12. Birkmeyer JD, Finks JF, O'Reilly A et al (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442. https://doi.org/10.1056/NEJMsa1300625
13. Fecso AB, Szasz P, Kerezov G, Grantcharov TP (2017) The effect of technical performance on patient outcomes in surgery: a systematic review. Ann Surg 265(3):492–501. https://doi.org/10.1097/SLA.0000000000001959
14. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:71. https://doi.org/10.1136/bmj.n71
15. Kelley WE Jr (2008) The evolution of laparoscopy and the revolution in surgery in the decade of the 1990s. JSLS 12(4):351–357
16. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C (2016) PRESS peer review of electronic search

strategies: 2015 guideline statement. J Clin Epidemiol 75:40–46. https://doi.org/10.1016/j.jclinepi.2016.01.021
17. Horsley T, Dingwall O (2011) Sampson M (2011) Checking reference lists to find additional studies for systematic reviews. Cochrane Database Syst Rev 2011(8):MR000026. https://doi.org/10.1002/14651858.MR000026.pub2
18. Peterson J, Welch V, Losos M, Tugwell P (2011) The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute, Ottawa
19. Wells G, Shea B, O'connell D et al (2011) The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa Hospital Research Institute. Oxford. Asp, Ottawa
20. Viale L, Allotey J, Cheong-See F et al (2015) Epilepsy in pregnancy and reproductive outcomes: a systematic review and meta-analysis. Lancet 386(10006):1845–1852. https://doi.org/10.1016/S0140-6736(15)00045-8
21. Sobhy S, Zamora J, Dharmarajah K et al (2016) Anaesthesia-related maternal mortality in low-income and middle-income countries: a systematic review and meta-analysis. Lancet Glob Health 4(5):e320–e327. https://doi.org/10.1016/S2214-109X(16)30003-1
22. Papola D, Ostuzzi G, Thabane L, Guyatt G, Barbui C (2018) Antipsychotic drug exposure and risk of fracture: a systematic review and meta-analysis of observational studies. Int Clin Psychopharmacol 33(4):181–196. https://doi.org/10.1097/YIC.0000000000000221
23. Wang B, An X, Shi X, Zhang JA (2017) Management of endocrine disease: suicide risk in patients with diabetes: a systematic review and meta-analysis. Eur J Endocrinol 177(4):R169–R181. https://doi.org/10.1530/EJE-16-0952
24. Visser A, Geboers B, Gouma DJ, Goslings JC, Ubbink DT (2015) Predictors of surgical complications: a systematic review. Surgery 158(1):58–65. https://doi.org/10.1016/j.surg.2015.01.012
25. Dettori JR (2011) Loss to follow-up. Evid Based Spine Care J 2(1):7–10. https://doi.org/10.1055/s-0030-1267080
26. Popay J, Roberts H, Sowden A, et al (2006) Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme Version. 1:b92.
27. Arvidsson D, Berndsen FH, Larsson LG et al (2005) Randomized clinical trial comparing 5-year recurrence rate after laparoscopic versus Shouldice repair of primary inguinal hernia. Br J Surg 92(9):1085–1091. https://doi.org/10.1002/bjs.5137
28. Mills JT, Hougen HY, Bitner D, Krupski TL, Schenkman NS (2017) Does robotic surgical simulator performance correlate with surgical skill? J Surg Educ 74(6):1052–1056. https://doi.org/10.1016/j.jsurg.2017.05.011
29. Brajcich BC, Stulberg JJ, Palis BE et al (2021) Association between surgical technical skill and long-term survival for colon cancer. JAMA Oncol 7(1):127–129. https://doi.org/10.1001/jamaoncol.2020.5462
30. Varban OA, Thumma JR, Finks JF, Carlin AM, Ghaferi AA, Dimick JB (2021) Evaluating the effect of surgical skill on outcomes for laparoscopic sleeve gastrectomy: a video-based study. Ann Surg 273(4):766–771. https://doi.org/10.1097/SLA.0000000000003385
31. Stulberg JJ, Huang R, Kreutzer L et al (2020) Association between surgeon technical skills and patient outcomes. JAMA Surg 19:19
32. Curtis NJ, Foster JD, Miskovic D et al (2020) Association of surgical skill assessment with clinical outcomes in cancer surgery. JAMA Surg 155(7):590–598. https://doi.org/10.1001/jamasurg.2020.1004
33. Fecso AB, Bhatti JA, Stotland PK, Quereshy FA, Grantcharov TP (2019) Technical performance as a predictor of clinical outcomes in laparoscopic gastric cancer surgery. Ann Surg 270(1):115–120. https://doi.org/10.1097/SLA.0000000000002741
34. Goldenberg MG, Goldenberg L, Grantcharov TP (2017) Surgeon performance predicts early continence after robot-assisted radical prostatectomy. J Endourol 31(9):858–863. https://doi.org/10.1089/end.2017.0284
35. Scally CP, Varban OA, Carlin AM, Birkmeyer JD, Dimick JB, Michigan Bariatric Surgery C. (2016) Video ratings of surgical skill and late outcomes of bariatric surgery. JAMA Surg 151(6):1160428. https://doi.org/10.1001/jamasurg.2016.0428
36. Paterson C, McLuckie S, Yew-Fung C, Tang B, Lang S, Nabi G (2016) Videotaping of surgical procedures and outcomes following extraperitoneal laparoscopic radical prostatectomy for clinically localized prostate cancer. J Surg Oncol 114(8):1016–1023. https://doi.org/10.1002/jso.24484
37. Hogg ME, Zenati M, Novak S et al (2016) Grading of surgeon technical performance predicts postoperative pancreatic fistula for pancreaticoduodenectomy independent of patient-related variables. Ann Surg 264(3):482–491. https://doi.org/10.1097/SLA.0000000000001862
38. Mackenzie H, Ni M, Miskovic D et al (2015) Clinical validity of consultant technical skills assessment in the English National Training Programme for Laparoscopic Colorectal Surgery. Br J Surg 102(8):991–997. https://doi.org/10.1002/bjs.9828
39. Madani A, Vassiliou MC, Watanabe Y et al (2017) What are the principles that guide behaviors in the operating room?: Creating a framework to define and measure performance. Ann Surg 265(2):255–267. https://doi.org/10.1097/SLA.0000000000001962
40. Varban OA, Sheetz KH, Cassidy RB et al (2017) Evaluating the effect of operative technique on leaks after laparoscopic sleeve gastrectomy: a case-control study. Surg Obes Relat Dis 13(4):560–567. https://doi.org/10.1016/j.soard.2016.11.027
41. Ponce J (2018) Impact of different surgical techniques on outcomes in laparoscopic sleeve gastrectomies: first report from the metabolic and bariatric surgery accreditation and quality improvement program (MBSAQIP). Ann Surg 267(3):e52. https://doi.org/10.1097/SLA.0000000000002076
42. Le AT, Huang B, Hnoosh D et al (2017) Effect of complications on oncologic outcomes after pancreaticoduodenectomy for pancreatic cancer. J Surg Res 214:1–8. https://doi.org/10.1016/j.jss.2017.02.036
43. Park EJ, Baik SH, Kang J et al (2016) The impact of postoperative complications on long-term oncologic outcomes after laparoscopic low anterior resection for rectal cancer. Medicine (Baltimore) 95(14):e3271. https://doi.org/10.1097/MD.0000000000003271
44. Beecher SM, O'Leary DP, McLaughlin R, Kerin MJ (2018) The impact of surgical complications on cancer recurrence rates: a literature review. Oncol Res Treat 41(7–8):478–482. https://doi.org/10.1159/000487510
45. Greenberg CC, Ghousseini HN, Pavuluri Quamme SR, Beasley HL, Wiegmann DA (2015) Surgical coaching for individual performance improvement. Ann Surg 261(1):32–34. https://doi.org/10.1097/SLA.0000000000000776
46. ABS to Explore Video-Based Assessment in Pilot Program Launching June 2021. The American Board of Surgery; 2021. Accessed 2021/4/22. https://www.absurgery.org/default.jsp?news_vba04.21
47. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M (2015) The impact of feedback of intraoperative technical performance in surgery: a systematic review. BMJ Open 5(6):e006759. https://doi.org/10.1136/bmjopen-2014-006759
48. Ritter EM, Gardner AK, Dunkin BJ, Schultz L, Pryor AD, Feldman L (2020) Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative

performance assessment. Surg Endosc 34(7):3176–3183. https://doi.org/10.1007/s00464-019-07089-y
49. Watanabe Y, Bilgic E, Lebedeva E et al (2016) A systematic review of performance assessment tools for laparoscopic cholecystectomy. Surg Endosc 30(3):832–844. https://doi.org/10.1007/s00464-015-4285-8
50. Bilgic E, Al Mahroos M, Landry T, Fried GM, Vassiliou MC, Feldman LS (2019) Assessment of surgical performance of laparoscopic benign hiatal surgery: a systematic review. Surg Endosc 33(11):3798–3805. https://doi.org/10.1007/s00464-019-06662-9
51. Tam V, Zeh HJ 3rd, Hogg ME (2017) Incorporating metrics of surgical proficiency into credentialing and privileging pathways. JAMA Surg 152(5):494–495. https://doi.org/10.1001/jamasurg.2017.0025
52. Deijen CL, Velthuis S, Tsai A et al (2016) COLOR III: a multicentre randomised clinical trial comparing transanal TME versus laparoscopic TME for mid and low rectal cancer. Surg Endosc 30(8):3210–3215. https://doi.org/10.1007/s00464-015-4615-x
53. Scott DJ, Rege RV, Bergen PC et al (2000) Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech A 10(4):183–190. https://doi.org/10.1089/109264200421559
54. Joosten M, Bokkerink GMJ, Verhoeven BH, Sutcliffe J, de Blaauw I, Botden S (2021) Are self-assessment and peer assessment of added value in training complex pediatric surgical skills? Eur J Pediatr Surg 31(1):25–33. https://doi.org/10.1055/s-0040-1715438
55. Scully RE, Deal SB, Clark MJ et al (2020) Concordance between expert and nonexpert ratings of condensed video-based trainee operative performance assessment. J Surg Educ 77(3):627–634. https://doi.org/10.1016/j.jsurg.2019.12.016
56. Deal SB, Lendvay TS, Haque MI et al (2016) Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. Am J Surg 211(2):398–404. https://doi.org/10.1016/j.amjsurg.2015.09.005
57. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D (2012) Rater training to support high-stakes simulation-based assessments. J Contin Educ Health Prof Fall 32(4):279–286. https://doi.org/10.1002/chp.21156
58. Bilgic E, Watanabe Y, McKendy K et al (2016) Reliable assessment of operative performance. Am J Surg 211(2):426–430. https://doi.org/10.1016/j.amjsurg.2015.10.008