Validity of video-based intraoperative self-assessment by surgical trainees

Saba Balvardi, MD

McGill University, Montreal

December 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Masters of Science – Epidemiology, Biostatistics and Occupational Health

TABLE OF CONTENTS

ABSTRACT

Introduction: 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. Furthermore, these recordings also offer new opportunities for trainees to extend technical learning outside of the operating room through self-assessment, in the context of well-documented reduction in intraoperative technical learning. Therefore, in this thesis we first aimed to systematically review the existing literature on the association between intraoperative technical performance, measured using VBAs, and patient outcomes. We then aimed to contribute evidence regarding the validity of intraoperative performance assessment tools for video-based self-assessment by general surgery trainees as a technical learning adjunct.

Methods: For the systematic review, 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 (≤30days) and/or long-term postoperative outcomes. Results were appraised descriptively as study heterogeneity precluded meta-analysis. For the validity assessment, general surgery trainees were recruited from McGill University and submitted recording of their performance in laparoscopic cholecystectomy. Operative performance was measured by the attending surgeon and trainees using global and hybrid (global + procedure-specific) assessment tools (GOALS and OPRS, respectively). The validity of GOALS and OPRS for trainee self-assessment was investigated by testing the hypotheses that self-assessment scores correlate with (H1) expert assessment scores, (H2) expert entrustability score, and (H3) procedure time; and that (H4) self-assessment based on these instruments differentiates junior (postgraduate year

(PGY)1-3) and senior trainees (PGY4-5), as well as (H5) simple (Visual Analogue Scale [VAS]≤ 4) versus complex cases (VAS>4).

Results: In our systematic review, 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). Better intraoperative performance was associated with fewer short term 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. In our validity study a total of 35 videos from 11 trainees were submitted (45% female and 45% senior trainees) and self-assessed. Our data supported 2 out of 5 hypotheses (H1 and H4) for the GOALS tool and 3 out of 5 hypotheses (H1, H4 and H5) for the OPRS.

Conclusion: Our results supported an association between superior intraoperative technical performance measured using surgical videos and improved short-term postoperative outcomes. Furthermore, we demonstrated stronger evidence supporting the validity of OPRS as a trainee video-based self-assessment tool, suggesting an advantage for assessments that include procedure specific items compared to global assessments alone. Given the reduced operative exposure of surgical trainees, and the association between technical proficiency and patient outcomes, strategies such as self-assessment to expand skills training outside the operating room are becoming crucial.

RÉSUMÉ

Introduction: Les efforts visant à améliorer la sécurité chirurgicaux ont traditionnellement accordé peu d'importance aux performances peropératoires, en partie dû aux obstacles de quantification. L'évaluation basée sur la vidéo (EBV) offre une opportunité d'évaluation en aveugle de la performance du chirurgien. Ces enregistrements offrent également de nouvelles opportunités pour les résidents d'étendre leur apprentissage technique à l'extérieur de la salle opératoire grâce à l'auto-évaluation dans un contexte de réduction bien documentée de l'apprentissage technique peropératoire. Par conséquent, dans cette thèse, nous avons d'abord cherché à passer en revue systématique la littérature existante sur l'association entre les performances techniques peropératoires, mesurées à l'aide des EBV, et les résultats pour les patients. Nous avons ensuite cherché à apporter des preuves concernant la validité des outils d'évaluation des performances peropératoires pour l'auto-évaluation par vidéo des résidents en chirurgie générale.

Méthodes: Les principales bases de données (Medline, Embase, Cochrane Database et Web of Science) ont été utilisées pour rechercher systématiquement les études évaluant l'association entre les performances techniques peropératoires mesurées par des outils étayés sur des preuves de validité et les résultats post-opératoires à court (≤ 30 jours) ou long terme. Les résultats ont été évalués de manière descriptive car l'hétérogénéité des études excluait la méta-analyse. Pour l'évaluation de la validité, des résidents en chirurgie générale ont été recrutés à l'Université McGill et ont soumis un enregistrement de leur performance en cholécystectomie laparoscopique. Les performances opératoires ont été mesurées par le chirurgien traitant et par les résidents à l'aide d'outils d'évaluation globaux et hybrides (GOALS

et OPRS, respectivement). La validité des outils GOALS et OPRS pour l'auto-évaluation des résidents a été étudiée en testant les hypothèses selon lesquelles les scores d'auto-évaluation sont en corrélation avec (H1) les scores d'évaluation des experts, (H2) le score de confiance des experts et (H3) la durée de la procédure ; et que l'auto-évaluation (H4) basée sur ces instruments différencie les résidents juniors (année de formation 1-3) des résidents seniors (année de formation 4-5), ainsi que (H5) les cas simples (Échelle Visuelle Analogique [EVA]≤ 4) des cas complexes (EVA>4).

Résultats: Dans notre revue systématique, un total de 11 études observationnelles ont été identifiées impliquant 8 interventions différentes en chirurgie intestinale (n=4), colorectale (n=4), urologique (n=2) et hépatobiliaire (n=1). Une meilleure performance peropératoire était associée à moins de complications postopératoires à court terme (6 études sur 7), de réopérations (3 études sur 4) et de réadmissions (1 étude sur 4). Les résultats à long terme ont été moins souvent étudiés avec des résultats mitigés. Dans notre étude de validité, un total de 35 vidéos de 11 stagiaires ont été soumises (45 % de femmes et 45 % de résidents seniors) et auto-évaluées. Nos données appuient 2 hypothèses sur 5 (H1 et H4) pour l'outil GOALS et 3 hypothèses sur 5 (H1, H4 et H5) pour l'OPRS.

Conclusion: Nos résultats ont confirmé une association entre les performances techniques peropératoires supérieures (mesurées à l'aide des EBV) et l'amélioration des résultats postopératoires à court terme. De plus, nous avons démontré des preuves plus tangibles étayant la validité de l'OPRS en tant qu'outil d'auto-évaluation vidéo des résidents, suggérant un avantage pour les évaluations qui incluent des éléments spécifiques à la procédure comparé aux évaluations globales seules. Compte tenu de l'exposition opératoire réduite des résidents

en chirurgie et de l'association entre les compétences techniques et les résultats pour les patients, des stratégies telles que l'auto-évaluation pour étendre la formation aux compétences deviennent cruciales.

ACKNOWLEDGMENTS

I would like to thank Dr. Schwartzman for his consistent guidance without which this thesis would have not been possible. His support has been incredibly valuable in every step of this endeavor. Furthermore, I would like to thank Dr. Liane Feldman my thesis co-supervisor and clinical mentor for her unwavering support and mentorship. Since the beginning of my medical training, I have had the privilege of learning from Dr. Feldman as an exemplary clinician, a scientist, and a leader. I will be forever grateful to her and the knowledge that she has imparted on me.

I would also like to thank Dr. Julio Fiore Jr. who I am incredibly indebted to for the researcher I am today. As a young medical student Dr. Fiore trained me in the scientific method and trusted me with opportunities that have helped me get to the place I am today professionally, and I am forever grateful for his trust, mentorship, and friendship. I am also grateful to all of the other members of the Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, especially Ms. Pepa Kaneva for their invaluable help and support.

Lastly, I would like to acknowledge my mom, Narjes, my dad, Assad and my sister, Sahar. I owe them the person and professional that I am today. Their unwavering love, support and sacrifice has made my academic ambitions a reality. Finally, I dedicate this work to my sister, Sahar, the person who has been with me through it all. She has been my biggest source of love and support when we have both been away from family.

FINANCIAL SUPPORT

Salary support for Saba Balvardi provided by the Fonds de la Recherche de Québec – Santé (FRQS) and the Canadian Institutes of Health Research (CIHR) over a 2-year period.

Society of University Surgeons STORZ Award and McGill University Department of Surgery Surgeon Scientist Program operating grants provided support for the implementation of the original randomized controlled trial that this study is based on.

CONTRIBUTION OF AUTHORS

Saba Balvardi is the primary author of this thesis. She has been the primary author of both manuscripts included in this study and was significantly involved in study design, data collection, data synthesis and drafting of the manuscripts. Dr. Kevin Schwartzman is the primary thesis supervisor and Dr. Liane Feldman is the thesis co-supervisor. Both Dr. Feldman and Dr Schwartzman had significant roles in design, conception and synthesis of the two manuscripts included in this thesis and they provided unparalleled guidance and mentorship throughout the synthesis of this master's thesis.

Contribution of each the co-authors for the two included manuscripts are outlined below.

Manuscript 1:

The association between video-based assessment of intraoperative technical performance and patient outcomes: A Systematic Review

Authors:

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

[1] *Department of Surgery, McGill University, Montreal, Quebec, Canada*

[2] *Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, Quebec, Canada*

[3] *Respiratory Division, Department of Medicine, McGill University and McGill International Tuberculosis Centre, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada*

- Study conception and design: SB, JFF, LSF
- Data acquisition: SB, AK, MH
- Analysis and interpretation of data: SB, CM, MV, LL, KS, JFF, LSF
- Drafting of manuscript: SB, KS, JFF, LSF
- Critical revision: SB, AK, MH, CM, MV, LL, KS, JFF, LSF

Manuscript 2:

Validity of Video-Based Intraoperative Self-Assessment by Surgical Trainees

Saba Balvardi, MD[1,2,3] Koorosh Semsar-Kazerooni, MS2[2] Pepa Kaneva, MSc[2] Carmen Mueller, MD Med[1,2] Melina Vassiliou, MD Med[1,2] Mohammed Al Mahroos, MD[1] Julio F Fiore Jr., PhD[1,2] Kevin Schwartzman, MD MPH[3] Liane S Feldman, MD[1,2]

[1] Department of Surgery, McGill University, Montreal, Quebec, Canada

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

[3] *Respiratory Division, Department of Medicine, McGill University and McGill International Tuberculosis Centre, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada*

- Study conception and design: SB, JFF, KS, LSF
- Data acquisition: SB, KSK, PK, MV, MA
- Analysis and interpretation of data: SB, JFF, KS, LSF
- Drafting of manuscript: SB, JFF, KS, LSF
- Critical revision: SB, KS, PK, CM, MV, MA, JFF, KS, LSF

CHAPTER 1: Introduction and Literature Review

1.1 Video-based assessment of intraoperative performance and its relationship with surgical outcome

Despite rapid advances in quality and outcomes, surgical care is highly complex and is associated with complications and other adverse events. Previous literature has reported that up to 54% of peri-operative adverse events originate in the operating room with up to one-third of these events being preventable.[1-3] Yet, efforts to improve surgical safety and outcomes have traditionally placed little emphasis on intraoperative performance while focusing on the system and process of care.[4-6] This approach disregards the principle that optimal outcomes can be more optimally achieved in well-designed systems where individual performance is also performed at the highest standard.[4] As a proof of principle, 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 short term rates of complications, reoperations and readmissions.[7, 8] Others have subsequently reported the same association in a wide variety of procedures including colectomy, pancreaticoduodenectomy, sleeve gastrectomy and prostectomy,[9-13] underlying the importance of measuring and improving intraoperative performance in surgical quality improvement.

Traditionally, one of the main obstacles to improving intraoperative performance in surgical outcomes has been difficulty in accessing information on 'what happens in the operating room'.[4, 5] This difficulty has been attributed to the interference caused by directly observing intraoperative performance in a high-stakes environment and the lack of appropriate tools for accurate and reliable assessment of intraoperative performance.[4, 5] Instead, indirect surrogate

measures have been used to reflect intraoperative performance such as surgeon years of experience[14, 15], fellowship training[16], surgeon case-volume[17], hospital case-volume[17] or level[18], or postoperative indicators considered to reflect intraoperative skills such as imaging findings[19] or pathology specimens[20]. While a systematic review by Fecso *et al.* of the literature up to 2014 on the effect of technical performance on patient outcomes in surgery identified 24 studies in seven surgical specialties, 20 of these relied on indirect surrogate measures of technical performance.[8]

There are inherent limitations associated with use of surrogate measures including confounding and information bias.[8, 21] Direct intraoperative performance assessment decreases these risks, and also directly identifies actionable targets for coaching and quality improvement.[8] The expansion of image guided surgery including laparoscopic, endoscopic and robotic operations facilitates capture, storage and sharing of recorded procedures. This allows for visual capture of the entirety or parts of intraoperative performance for storage and assessment at a later time.

Video-based assessment (VBA) of image-guided procedures provides a valuable opportunity to measure intraoperative performance directly while minimizing observer interference and bias related to unblinded in-theater evaluations.[22, 23] These benefits suggest multiple diverse applications for VBA in surgical education, quality improvement and even credentialling. VBA can be used for 'formative' assessment aimed to track progress of the operator and identifying need for focused training or coaching for technical improvement. [4, 6, 24] VBAs can also provide a more high-stakes 'summative' assessment for informing competency decisions for certification or credentialing purposes.[23, 25, 26] The Michigan Bariatric Surgery Collaborative is an example of a surgical quality assurance initiative that uses VBA in a formative assessment role for

identifying bariatric surgeons in this network who can benefit from technical coaching for improving technical outcomes.[23] The LapCo program in the UK required a summative VBA to ensure competency for practicing surgeons transitioning from open to laparoscopic colectomy surgery (Hanna et al Ann Surg 2020). The Surgical Skill Qualification System in Japan is using VBAs for summative assessment for licencing of laparoscopic surgeons aimed at setting standards for surgeons certified to perform laparoscopic surgery.[26]

While there is enormous potential for the use of VBA in quality improvement, this is still an actively developing field. VBAs have many different acquisition, storage, and measurement characteristics that can influence the accuracy and reliability of the intraoperative assessment.[8] These include the nature of the submitted intraoperative recording (edited versus unedited)[27, 28], rater qualifications (expert versus peer versus crowdsourcing) [23, 29-31], approach to rater training and the type and validity of assessment tools (generic vs. procedure specific).[32, 33] These features have been sparsely studied with conflicting conclusions.[8] Standards based on expert opinion have been suggested; for example, Fecso *et al*. advocate for evaluation of unedited intraoperative recordings by expert assessors who have been trained in the use of validated procedure-specific assessment tools for summative assessment.[8] However higher level evidence is required prior to adoption of these recommendations.

1.2 Intraoperative Assessment Tools

In surgical training and continuing education, efforts to objectively and systematically assess intraoperative technical proficiency have fallen well behind assessment of knowledge and judgement.[34] Over the last decades, intraoperative assessment tools have been developed for

use in a range of settings and procedures, allowing for more systematic, reproducible, and objective assessment of intraoperative skills.[32, 34] These assessment tools can be broadly classified into three categories: global generic skills assessment tools, procedure-specific assessment tools and hybrid assessment tools.[35]

Global assessment tools evaluate surgical techniques that are generic and applicable to any procedure.[32] For example, the Global Operative Assessment of Laparoscopic Skills (GOALS) is a generic assessment tool that includes five domains important in laparoscopic surgery including depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy of the operator.[32] Conversely, procedure specific assessment tools include items that allow for assessment of skill in each step of a given procedure.[36] For example, in a procedure specific assessment tool for laparoscopic cholecystectomy, one item would assess 'adequacy of dissection of cystic artery' with rating anchors focusing on lack of arterial bleeding during the dissection.[37] Hybrid assessment tools combine both global and procedure specific items. Each type of assessment can offer advantages in different contexts.[35, 36] Global assessment tools may have advantages in the formative assessment of trainees when a given procedure is divided between multiple residents (i.e., in laparoscopic cholecystectomy the senior trainee dissecting the triangle of Calot while the junior trainee dissects the gallbladder bed) or if proficiency in laparoscopic surgery is being followed over time across various procedures requiring similar skills.[35] Alternatively, procedure specific instruments enable more specific feedback to trainees and define more specific targets for improvement.[35]

It should also be highlighted that the choice of assessment tool will vary depending on the context and purpose of the evaluation. These tools can be used for direct intraoperative

assessment (i.e., by a supervising attending surgeon in the context of residency training) or video-based assessment. Moreover, the purpose of the intraoperative assessment tool can be either formative or summative. In formative assessment the instrument is used to systematically identify areas where the trainee is in need of improvement or the attending surgeon will require coaching to achieve set standards.[34, 38] Summative assessment is used for purposes of high stake evaluation in credentialing or licencing settings.[38]

Evidence supporting the validity of an assessment tool should also take into consideration the intended context and purpose of assessment.[35] Under the contemporary framework, a validation study aims to contribute evidence to support the interpretation of the assessment results under a specific set of conditions; validity is not considered an absolute property of the assessment tool itself.[39] For example, if an assessment tool has evidence to support its validity for use under a given set of conditions, it might not be appropriate under a different set of conditions or for a different purpose without additional evidence supporting its new intended use.[35]

Most of the currently available intraoperative assessment tools were developed and validated to evaluate surgical trainees and focus on psychomotor skills.[5, 23, 40] These skills are paramount for safety. However, they fail to encompass all the domains of surgical expertise such as advanced cognitive skills and decision making, and have important limitations if the intended use is for high-stakes summative assessment (i.e., credentialing).[40] It is therefore not surprising that in bariatric surgery, VBA scores using the generic Objective Structured Assessment of Technical Skill (OSATS) tool correlated with early complications (i.e., safety) but not long term outcomes (i.e., effectiveness).[13] OSATS was originally developed to evaluate trainee

performance for basic surgical skills in a box trainer. [41] This is an example of how evidence of validity for a tool used for formative assessment of trainee skills which are important for operative safety (i.e., respect for tissue, and operation flow) may not guarantee its validity for use in another context and purpose.[41, 42]

Consequently, the context in which validity evidence was collected is of paramount importance to inform selection of the most appropriate tool.[35, 43] Laparoscopic cholecystectomy is a commonly performed general surgical procedure which can be delegated to trainees when they demonstrate appropriate competencies, and is therefore a useful model in surgical education. Watanabe *et al.* performed a systematic review to identify assessment tools that have been used for this procedure and to summarize the context under which validity evidence was available.[35] This review identified the GOALS (a global assessment) and Operative Performance Rating System (OPRS; a hybrid assessment) as two assessment tools with existing validity evidence supporting their use for direct and video-based formative and summative assessment of surgical trainees by attending surgeons. However, there was limited evidence supporting their use as video based formative self-assessment tools.[35] Self-assessment will be reviewed in further detail in section 1.4.

1.3 Role of Video-based assessment in surgical training and continuing education

Restriction of working hours during residency has reduced the operative exposure of surgical trainees, and almost one third of surgical graduates do not feel confident in their ability to perform certain procedures independently.[44] Other modern challenges to surgical education includes the COVID-19 pandemic that has further reduced exposure for trainees to elective

surgery.[45] This was the result of strategies used to expand healthcare system capacity to treat patients with COVID-19 including mandated cessation of non-essential surgical procedures to re-deploy staff and liberate hospital beds.[46] Hence, enhancement of training both inside and outside the operating room has become crucial.[47] While the acquisition of fundamental psychomotor skills is enabled through simulation, expertise requires higher level cognitive skills.[40] Video-recording of surgical procedures has offered new opportunities to trainees to extend technical learning to outside the operating room. Consequently, this has led to growing interest in the potential use of video to augment traditional surgical training.[47]

A recent systematic review by Green *et al.* synthesized the evidence for the uses of intraoperative video recording in surgical education.[47] This review identified 19 studies, mostly in the setting of laparoscopic or arthroscopic surgery. Videos were viewed for supplementation of preoperative (11 studies) or postoperative (9 studies) education.[47] Seven studies assessed theoretical knowledge acquisition and 16 studies investigated technical skill acquisition. Compared to nonvideo training groups, 13 out of 19 studies demonstrated significant improvement in knowledge and 15 out of 19 studies showed improvement in technical skills in trainees with receiving video-based educational interventions.[47] This review demonstrated that postoperative operative recording review and feedback was the most effective video review modality.[47] While this review and other more recent studies have demonstrated that video-assisted postoperative structured feedback by expert surgeons significantly improves laparoscopic skill acquisition in surgical trainees,[47-49] the reported median time commitments for delivering this intervention for each resident for one procedure is between 40-50 minutes.[48]

Hence, this method can prove to be quite resource intensive and has limited feasibility outside of research settings.

1.4 Role of self-assessment in surgical training

Self-assessment can be defined as a self-driven process aimed towards ongoing self-improvement.[50] Self-assessment is an integral part of medical learning that encourages improvement, life-long learning and self-regulation.[51] Guided self-assessment has been successfully demonstrated to improve performance in medical and non-medical fields such as sports and music.[52, 53] Self-assessment is therefore a pertinent skill to develop as it acts as a self-regulated educational tool especially in surgical training where external feedback is not always readily available, particularly after formal training ends.[51]

A systematic review of self-assessment in technical tasks in surgery by Zevin *et al.* reported mixed results regarding the accuracy of trainee self-assessment with the majority of the studies reporting higher self-assessment scores compared to expert assessment.[51] These findings have been partly attributed to methodological limitations of previous studies, including the use of unvalidated assessment tools and recall bias (i.e. poor recall of intraoperative events by trainees after the fact).[51]

Self-assessment can be affected by intrinsic or demographic factors and cognitive factors. For example, a meta-analysis of self-assessment in medical students reported that female trainees tended to underestimate their performance compared to male students and that the accuracy of self-assessment increased with trainee experience.[54] Cognitive factors such as 'memory bias' have also been reported to affect accuracy of self-reflection. Memory bias is a defense mechanism that encourages poor recall of personal failures, in order to decrease unhappiness and

despair.[50] Increased experience, use of video review and use of valid and reliable assessment tools with unambiguous behavioral anchors were associated with improved accuracy of self-assessment.[51, 53] Video-based self-reflection can readily address recall bias and memory bias, and valid assessment tools with clear performance anchors can address the lack of accuracy and inconsistency in interpretation of items.[51, 55] These interventions have therefore increased interest in the use of self-assessment as an educational tool in procedural learning.

CHAPTER 2: Thesis Objectives

The objectives of our study, divided in two parts, are the following:

Part 1 – To systematically review and summarize the existing literature on the association between intraoperative technical performance via VBA and patient outcomes in practicing surgeons (P: Surgeons in practice, I: VBA of Intraoperative performance, C: No specific intervention, O: Postoperative outcome)

Part 2 – To generate evidence for the validity of two intraoperative assessment tools for formative video-based self-assessment by general surgery trainees in laparoscopic cholecystectomy.

CHAPTER 3: The association between video-based assessment of intraoperative technical performance and patient outcomes: A Systematic Review

3.1 Preamble to Manuscript 1

Conventional wisdom assumes that a surgeon's skill in the operating room affects patient outcomes. However, most efforts to improve surgical safety and outcomes focus on perioperative care with very little emphasis on measuring and improving operative performance.[4] Some of the barriers that have hindered this field of research include difficulty in gathering information in the operating room and lack of valid and reliable tools for assessment of intra-operative performance.[4, 5] Evidence suggests an association between surgical skill and patient outcomes.[7, 8] Yet previous studies mainly relied on indirect measures of skill such as postoperative imaging or pathological specimens rather than measurement of the performance of the operation itself.[8]

The expansion of image guided surgery including laparoscopic and robotic surgery allows for easy capture, storage and sharing of recorded procedures. Video-based assessment (VBA) of recorded operative procedures provides a new opportunity to measure surgeon performance while minimizing barriers and biases related to direct in-theater evaluations.[22] Furthermore, recent development of global and procedure-specific assessment tools may enable more accurate measurement of intraoperative performance.[32] Consequently, the relationship between intra-operative technical skill and patient outcome has become an active area of

research. This new body of evidence needs to be formally synthesized to inform appropriate integration of VBA into credentialing, certification, coaching and quality improvement processes and identify gaps for future research.[7, 9] Therefore the objective of manuscript 1 was to systematically review and summarize the existing literature on the association between intraoperative technical performance measured using VBAs and patient outcomes.

MANUSCRIPT 1: The association between video-based assessment of intraoperative technical performance and patient outcomes: A Systematic Review

Authors:

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

[1] *Department of Surgery, McGill University, Montreal, Quebec, Canada*

[2] *Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, Quebec, Canada*

[3] Respiratory Division, Department of Medicine, McGill University and McGill International Tuberculosis Centre, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada

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

Running Head: technical performance and patient outcomes

Accepted for publication in Surgical Endoscopy, February 2022

3.2 ABSTRACT

Background: Efforts to improve surgical safety and outcomes have traditional 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 (≤30days) 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 10242. Short-term outcomes were reported in 8 studies (i.e., morbidity, mortality, 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

3.3 INTRODUCTION

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 while minimizing observer bias related to unblinded in-theater evaluations.[22, 23]

There is significant interest in the use of VBA of intraoperative performance for formative assessment in education and coaching.[4, 6, 24] 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.[25, 26] However, the use of VBA to inform competency decisions 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.[7] Yet there remain important limitations related to lack of standardized assessment tools and reliance on indirect observations of technical performance such as postoperative imaging or pathological specimen quality.[8] 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. 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.

3.4 MATERIALS and METHODS

This review was conducted and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA).[56] 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)[57] 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 (ie, 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), Web of Science (Thomson Reuters). The search strategies (**eMethods 1**) were developed by an experienced medical librarian according to best practice recommendations.[58] The reference list of the selected studies was screened for further studies that met the inclusion criteria. [59] 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)[60]. Any discrepancies were resolved by consensus between the reviewers or by consulting a third independent reviewer (LF). NOS is a validated system developed for 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.[61] 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. [62-65] We followed *a priori* criteria for risk of bias analysis based on the NOS guidelines, as outlined in Supplemental Digital Content 1.[66 67]

Data Synthesis

This systematic review was reported using a narrative synthesis approach.[68] Meta-analysis was precluded as the identified studies were heterogeneous with respect to population, exposure and outcome measures.

3.5 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 (Figure). There were 3 additional studies identified through other sources (cross referencing [n=2][69, 70] or expert suggestions of recent papers which had not yet been indexed in Medline[n=1][71]. Twenty-three articles were excluded (articles and reasons for exclusion are listed in Supplemental Digital Content 1) and 11 articles met eligibility criteria.[7, 9-13, 71-75]

Characteristics of the included studies are summarized in Table 1. All were observational studies (10 cohort and 1 case-control study) published after the landmark paper by Birkmeyer *et al*.[7] Eight of 11 studies were multicenter collaborations. Two studies involved urologic procedures[10, 75] with the remainder involving general surgery procedures (foregut/bariatric[n=4], colorectal[n=4] and hepatobiliary surgery[n=1]).[7, 9, 11-13, 71-74] 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 to 10242 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.[8] 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. 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. Five studies involved VBA of the entire procedure. In 10 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 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 5 studies, participating surgeons chose and submitted one video as representative of their 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 6 studies, videos were available for each case and the association between intraoperative technical performance and outcomes were analyzed for each patient.

Quality assessment of each study was performed using the NOS tool.[61] 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)[7, 9, 11, 13, 71-74] followed by bias in measurement of exposure and non-disclosure of frequency and handling of missing data (n=10)[7, 9, 11-13, 71-75] . 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 (≤30days) or long-term (>30days). Short-term outcomes 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. Of the

3 studies with low risk of bias,[7, 13, 72] 2 demonstrated an association between better intraoperative performance and fewer postoperative complications (rate reduction between 9.2% and 5.1%)[7, 72]. Better intraoperative performance was associated with fewer reoperations in 3 of 4 studies (rate reduction between 0.7%-2.5%), including all 3 studies with low risk of bias.[7, 12, 13, 72]Better intraoperative performance had an association with fewer readmission in only 1 of 4 studies;[7] only one of these studies (that showed no association) had a high risk of bias.[72] All studies looking at ED visits and mortality were of low risk of bias.[7, 13, 72] One of 2 studies showed an association between better intraoperative performance and lower ED visits and mortality.[7]

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)[13, 74], and patient satisfaction (1 of 1 study with high risk of bias)[74], but not cancer recurrence (0 of 1 study with high risk of bias)[12]. 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[71] with a second study with high risk of bias reporting a large but non-statistically significance increase in overall survival. [12] 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[10] and one with high risk of bias[75]) (Table 4). Four studies reported the association between intraoperative technical performance and pathological outcomes. [11, 12, 71, 75] Of the 3 studies investigating the association between intraoperative technical performance and lymph node yield, 2 showed no association[12, 71] and 1 showed a significant association (13

vs. 18 LNs in colon cancer).[25] 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 ≥1 mm and distal margin ≥1 mm) [12] and another reported an association with the distal margin in left colon cancer surgery (median 3 vs. 4 cm).[11]

3.6 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 included only one study [7] where an intraoperative assessment 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.[8] Our systematic review was further strengthened with compliance with

PRISMA methodological standards and the use of cross-referencing to maximize our literature search.[56, 59]

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, 40] 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 (eg. oversewing versus not oversewing of the staple-line or length of the roux-limb in bariatric surgery) may also be associated with postoperative outcomes.[76, 77] 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.[13] 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. [78-80]

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.[6, 81] Similarly, there is interest in using VBA to measure and improve surgical techniques from leading groups such as the American Board of Surgery.[82] 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 [83], 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, 33] There is limited evidence supporting the use of the generic assessment tools identified in this review for summative video-based evaluation in practicing surgeons .[35,43] 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. [25,32] This work is critical as automated metrics of performance using computer vision and machine learning are rapidly being developed.[84] 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.[85]

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 published recommendations for minimizing measurement error when using VBAs.[8, 28] 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.[33]

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 with familiarity in using or developing intraoperative assessment tools. Use of multiple peer raters (as opposed to 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.[23] The literature supporting peer VBA assessment in comparison to expert assessment (the default gold standard) has been mixed [29, 30] with supporting evidence for their use in evaluating simple tasks such as knot tying [86] and in the presence of added information such as intraoperative audio[30]. To our knowledge, no studies support the use of peer assessment using only visual feedback from VBA in complex procedures. Until more evidence is available for optimizing the accuracy of peer assessment, use of expert assessment should be prioritized in future studies.

There was also wide range of definitions for rater training, ranging from passive training based on descriptive manuals[12] to full training programs with continuous calibration of the

assessors.[73] Only one of 5 studies that used peer assessors described any attempts at rater training. Lack of familiarity with the nuances of assessment tools can result in non-differential measurement error. This results in underestimation of the effect size and biases the analysis towards 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.[8, 23, 87]

Inconsistency in the association between intraoperative technical performance and outcomes between studies may be related to other issues in study design. Almost half of the studies utilized a single submitted video chosen by the participating surgeon. This 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. However, surgeons would likely select their "best" videos which would bias the results towards 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.[88]

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 their surgeon versus a consecutive cohort of patients where video and outcome data were both available.

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 greater intraoperative technical performance and lower 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.

DISCLOSURES:

Saba Balvardi: no conflict of interest

Anitha Kammili: no conflict of interest

Melissa Hanson: no conflict of interest

Carmen Mueller: no conflict of interest

Melina Vassiliou: no conflict of interest

Lawrence Lee: no conflict of interest

Julio F Fiore Jr.: no conflict of interest

Liane S Feldman: no conflict of interest

3.7 REFERENCES

1. Zegers M, de Bruijne MC, Wagner C, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. *Qual Saf Health Care*. Aug 2009;18(4):297-302. doi:10.1136/qshc.2007.025924
2. Kable AK, Gibberd RW, Spigelman AD. Adverse events in surgical patients in Australia. *Int J Qual Health Care*. Aug 2002;14(4):269-76. doi:10.1093/intqhc/14.4.269
3. Fabri PJ, Zayas-Castro JL. Human error, not communication and systems, underlies surgical complications. *Surgery*. Oct 2008;144(4):557-63; discussion 563-5. doi:10.1016/j.surg.2008.06.011
4. Dimick JB, Varban OA. Surgical video analysis: an emerging tool for improving surgeon performance. *BMJ Qual Saf*. Aug 2015;24(8):490-1. doi:10.1136/bmjqs-2015-004439
5. Feldman LS, Pryor AD, Gardner AK, et al. SAGES Video-Based Assessment (VBA) program: a vision for life-long learning for surgeons. *Surg Endosc*. Aug 2020;34(8):3285-3288. doi:10.1007/s00464-020-07628-y
6. Yanes AF, McElroy LM, Abecassis ZA, Holl J, Woods D, Ladner DP. Observation for assessment of clinician performance: a narrative review. *BMJ Qual Saf*. Jan 2016;25(1):46-55. doi:10.1136/bmjqs-2015-004171
7. Bilgic E, Valanci-Aroesty S, Fried GM. Video Assessment of Surgeons and Surgery. *Adv Surg*. Sep 2020;54:205-214. doi:10.1016/j.yasu.2020.03.002
8. Greenberg CC, Dombrowski J, Dimick JB. Video-Based Surgical Coaching: An Emerging Approach to Performance Improvement. *JAMA Surg*. Mar 2016;151(3):282-3. doi:10.1001/jamasurg.2015.4442
9. Grenda TR, Pradarelli JC, Dimick JB. Using Surgical Video to Improve Technique and Skill. *Ann Surg*. Jul 2016;264(1):32-3. doi:10.1097/SLA.0000000000001592
10. Mackenzie H, Cuming T, Miskovic D, et al. Design, delivery, and validation of a trainer curriculum for the national laparoscopic colorectal training program in England. *Ann Surg*. Jan 2015;261(1):149-56. doi:10.1097/SLA.0000000000000437
11. Mori T, Kimura T, Kitajima M. Skill accreditation system for laparoscopic gastroenterologic surgeons in Japan. *Minim Invasive Ther Allied Technol*. 2010;19(1):18-23. doi:10.3109/13645700903492969
12. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. Oct 10 2013;369(15):1434-42. doi:10.1056/NEJMsa1300625
13. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The Effect of Technical Performance on Patient Outcomes in Surgery: A Systematic Review. *Ann Surg*. Mar 2017;265(3):492-501. doi:10.1097/SLA.0000000000001959
14. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. Mar 29 2021;372:n71. doi:10.1136/bmj.n71
15. Kelley WE, Jr. The evolution of laparoscopy and the revolution in surgery in the decade of the 1990s. *JSLS*. Oct-Dec 2008;12(4):351-7.
16. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol*. Jul 2016;75:40-6. doi:10.1016/j.jclinepi.2016.01.021

17. Horsley T, Dingwall O, Sampson M. Checking reference lists to find additional studies for systematic reviews. *Cochrane Database Syst Rev*. Aug 10 2011;(8):MR000026. doi:10.1002/14651858.MR000026.pub2
18. Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. *Ottawa: Ottawa Hospital Research Institute*. 2011;
19. Wells G, Shea B, O'connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute. oxford. asp; 2011.
20. Viale L, Allotey J, Cheong-See F, et al. Epilepsy in pregnancy and reproductive outcomes: a systematic review and meta-analysis. *Lancet*. Nov 7 2015;386(10006):1845-52. doi:10.1016/S0140-6736(15)00045-8
21. Sobhy S, Zamora J, Dharmarajah K, et al. Anaesthesia-related maternal mortality in low-income and middle-income countries: a systematic review and meta-analysis. *Lancet Glob Health*. May 2016;4(5):e320-7. doi:10.1016/S2214-109X(16)30003-1
22. Papola D, Ostuzzi G, Thabane L, Guyatt G, Barbui C. Antipsychotic drug exposure and risk of fracture: a systematic review and meta-analysis of observational studies. *Int Clin Psychopharmacol*. Jul 2018;33(4):181-196. doi:10.1097/YIC.0000000000000221
23. Wang B, An X, Shi X, Zhang JA. MANAGEMENT OF ENDOCRINE DISEASE: Suicide risk in patients with diabetes: a systematic review and meta-analysis. *Eur J Endocrinol*. Oct 2017;177(4):R169-R181. doi:10.1530/EJE-16-0952
24. Visser A, Geboers B, Gouma DJ, Goslings JC, Ubbink DT. Predictors of surgical complications: A systematic review. *Surgery*. Jul 2015;158(1):58-65. doi:10.1016/j.surg.2015.01.012
25. Dettori JR. Loss to follow-up. *Evid Based Spine Care J*. Feb 2011;2(1):7-10. doi:10.1055/s-0030-1267080
26. Popay J, Roberts H, Sowden A, et al. Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version*. 2006;1:b92.
27. Arvidsson D, Berndsen FH, Larsson LG, et al. Randomized clinical trial comparing 5-year recurrence rate after laparoscopic versus Shouldice repair of primary inguinal hernia. *Br J Surg*. Sep 2005;92(9):1085-91. doi:10.1002/bjs.5137
28. Mills JT, Hougen HY, Bitner D, Krupski TL, Schenkman NS. Does Robotic Surgical Simulator Performance Correlate With Surgical Skill? *J Surg Educ*. Nov - Dec 2017;74(6):1052-1056. doi:10.1016/j.jsurg.2017.05.011
29. Brajcich BC, Stulberg JJ, Palis BE, et al. Association Between Surgical Technical Skill and Long-term Survival for Colon Cancer. *JAMA Oncol*. Jan 1 2021;7(1):127-129. doi:10.1001/jamaoncol.2020.5462
30. Varban OA, Thumma JR, Finks JF, Carlin AM, Ghaferi AA, Dimick JB. Evaluating the Effect of Surgical Skill on Outcomes for Laparoscopic Sleeve Gastrectomy: A Video-based Study. *Ann Surg*. Apr 1 2021;273(4):766-771. doi:10.1097/SLA.0000000000003385
31. Stulberg JJ, Huang R, Kreutzer L, et al. Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surgery*. 2020;19:19.

32. Curtis NJ, Foster JD, Miskovic D, et al. Association of Surgical Skill Assessment With Clinical Outcomes in Cancer Surgery. *JAMA Surg*. Jul 1 2020;155(7):590-598. doi:10.1001/jamasurg.2020.1004
33. Fecso AB, Bhatti JA, Stotland PK, Quereshy FA, Grantcharov TP. Technical Performance as a Predictor of Clinical Outcomes in Laparoscopic Gastric Cancer Surgery. *Ann Surg*. Jul 2019;270(1):115-120. doi:10.1097/SLA.0000000000002741
34. Goldenberg MG, Goldenberg L, Grantcharov TP. Surgeon Performance Predicts Early Continence After Robot-Assisted Radical Prostatectomy. *J Endourol*. Sep 2017;31(9):858-863. doi:10.1089/end.2017.0284
35. Scally CP, Varban OA, Carlin AM, Birkmeyer JD, Dimick JB, Michigan Bariatric Surgery C. Video Ratings of Surgical Skill and Late Outcomes of Bariatric Surgery. *JAMA Surg*. Jun 15 2016;151(6):e160428. doi:10.1001/jamasurg.2016.0428
36. Paterson C, McLuckie S, Yew-Fung C, Tang B, Lang S, Nabi G. Videotaping of surgical procedures and outcomes following extraperitoneal laparoscopic radical prostatectomy for clinically localized prostate cancer. *J Surg Oncol*. Dec 2016;114(8):1016-1023. doi:10.1002/jso.24484
37. Hogg ME, Zenati M, Novak S, et al. Grading of Surgeon Technical Performance Predicts Postoperative Pancreatic Fistula for Pancreaticoduodenectomy Independent of Patient-related Variables. *Ann Surg*. Sep 2016;264(3):482-91. doi:10.1097/SLA.0000000000001862
38. Mackenzie H, Ni M, Miskovic D, et al. Clinical validity of consultant technical skills assessment in the English National Training Programme for Laparoscopic Colorectal Surgery. *Br J Surg*. Jul 2015;102(8):991-7. doi:10.1002/bjs.9828
39. Madani A, Vassiliou MC, Watanabe Y, et al. What Are the Principles That Guide Behaviors in the Operating Room?: Creating a Framework to Define and Measure Performance. *Ann Surg*. Feb 2017;265(2):255-267. doi:10.1097/SLA.0000000000001962
40. Varban OA, Sheetz KH, Cassidy RB, et al. Evaluating the effect of operative technique on leaks after laparoscopic sleeve gastrectomy: a case-control study. *Surg Obes Relat Dis*. Apr 2017;13(4):560-567. doi:10.1016/j.soard.2016.11.027
41. Ponce J. 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*. Mar 2018;267(3):e52. doi:10.1097/SLA.0000000000002076
42. Le AT, Huang B, Hnoosh D, et al. Effect of complications on oncologic outcomes after pancreaticoduodenectomy for pancreatic cancer. *J Surg Res*. Jun 15 2017;214:1-8. doi:10.1016/j.jss.2017.02.036
43. Park EJ, Baik SH, Kang J, et al. The Impact of Postoperative Complications on Long-term Oncologic Outcomes After Laparoscopic Low Anterior Resection for Rectal Cancer. *Medicine (Baltimore)*. Apr 2016;95(14):e3271. doi:10.1097/MD.0000000000003271
44. Beecher SM, O'Leary DP, McLaughlin R, Kerin MJ. The Impact of Surgical Complications on Cancer Recurrence Rates: A Literature Review. *Oncol Res Treat*. 2018;41(7-8):478-482. doi:10.1159/000487510
45. Greenberg CC, Ghousseini HN, Pavuluri Quamme SR, Beasley HL, Wiegmann DA. Surgical coaching for individual performance improvement. *Ann Surg*. Jan 2015;261(1):32-4. doi: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. The impact of feedback of intraoperative technical performance in surgery: a systematic review. *BMJ Open*. Jun 15 2015;5(6):e006759. doi:10.1136/bmjopen-2014-006759
48. Ritter EM, Gardner AK, Dunkin BJ, Schultz L, Pryor AD, Feldman L. Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative performance assessment. *Surg Endosc*. Jul 2020;34(7):3176-3183. doi:10.1007/s00464-019-07089-y
49. Watanabe Y, Bilgic E, Lebedeva E, et al. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surg Endosc*. Mar 2016;30(3):832-44. doi:10.1007/s00464-015-4285-8
50. Bilgic E, Al Mahroos M, Landry T, Fried GM, Vassiliou MC, Feldman LS. Assessment of surgical performance of laparoscopic benign hiatal surgery: a systematic review. *Surg Endosc*. Nov 2019;33(11):3798-3805. doi:10.1007/s00464-019-06662-9
51. Tam V, Zeh HJ, 3rd, Hogg ME. Incorporating Metrics of Surgical Proficiency Into Credentialing and Privileging Pathways. *JAMA Surg*. May 1 2017;152(5):494-495. doi:10.1001/jamasurg.2017.0025
52. Deijen CL, Velthuis S, Tsai A, et al. COLOR III: a multicentre randomised clinical trial comparing transanal TME versus laparoscopic TME for mid and low rectal cancer. *Surg Endosc*. Aug 2016;30(8):3210-5. doi:10.1007/s00464-015-4615-x
53. Scott DJ, Rege RV, Bergen PC, et al. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. *J Laparoendosc Adv Surg Tech A*. Aug 2000;10(4):183-90. doi:10.1089/109264200421559
54. Joosten M, Bokkerink GMJ, Verhoeven BH, Sutcliffe J, de Blaauw I, Botden S. Are Self-Assessment and Peer Assessment of Added Value in Training Complex Pediatric Surgical Skills? *Eur J Pediatr Surg*. Feb 2021;31(1):25-33. doi:10.1055/s-0040-1715438
55. Scully RE, Deal SB, Clark MJ, et al. Concordance Between Expert and Nonexpert Ratings of Condensed Video-Based Trainee Operative Performance Assessment. *J Surg Educ*. May - Jun 2020;77(3):627-634. doi:10.1016/j.jsurg.2019.12.016
56. Deal SB, Lendvay TS, Haque MI, et al. Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. *Am J Surg*. Feb 2016;211(2):398-404. doi:10.1016/j.amjsurg.2015.09.005
57. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. Fall 2012;32(4):279-86. doi:10.1002/chp.21156
58. Bilgic E, Watanabe Y, McKendy K, et al. Reliable assessment of operative performance. *Am J Surg*. Feb 2016;211(2):426-30. doi:10.1016/j.amjsurg.2015.10.008

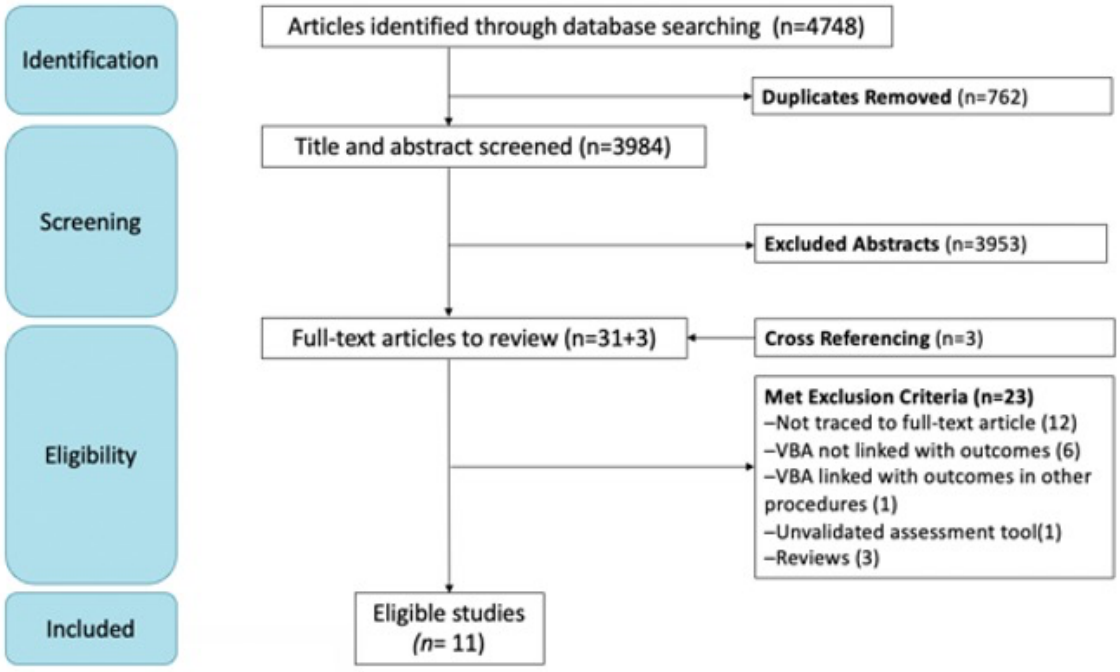


Figure 1: PRISMA flow diagram. (PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-analyses)[56].

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. [13]	2021	Multicenter Retrospective Cohort	United States	25 (35%)	3502	GS	Lap Sleeve Gastrectomy	Complications (30 d)	Readmission (30 d) Reoperation (30 d) ED visits (30 d) EBWL % (1 year)
Brajcich et al.[71]	2020	Multicenter Retrospective Cohort	United States	15 (NS)	609	GS	Lap Right Hemicolectomy	Survival (5-year)	Nil
Stulberg et al. [72]	2020	Multicenter Prospective Cohort	United States	17 (NS)	1120	GS	Lap Right Hemicolectomy	Complications (30 d) Mortality (30 d) Readmission (30 d) Reoperation (30 d)	Skill-related morbidity (30 d) [b] Skill-unrelated morbidity (30 d) [b]
Curtis et al.[12]	2020	Multicenter Prospective Cohort	Australia, New Zealand and United Kingdom	34 (100%)	176	GS	Lap Total Mesorectal Excision	Complications (30 d) Reoperation (30 d) Readmission (30 d)	Overall Survival (2-4 years) Cancer Recurrence (2-4 years)
Fecso et al. [73]	2019	Multicenter Retrospective Cohort	Canada	3 (10%)	61	GS	Lap Gastrectomy	Complications (30 d)	Nil
Goldenberg et al. [10]	2017	Single center Prospective Case-Control	Canada	1 (100%)	47	Urology	Robotic Assisted Radical Prostectomy	Continence (3 mo)	Nil
Scally et al. [74]	2016	Multicenter Retrospective Cohort	United States	20 (27%)	3631	GS	Lap Gastric Bypass	EBWL % (1 year)	Patient Satisfaction (1 year)
Paterson et al. [75]	2016	Single center Prospective Cohort	Scotland	1 (100%)	200	Urology	Extraperitoneal Lap Prostectomy	Continence (3 mo)	Continence (12 mo) Readmission (30, 90 and 120 d) [c] Reoperation (30, 90 and 120 d) [c] ED visits (30, 90 and 120 d) [c] Complications (30, 90 and 120 d) [c] Erectile dysfunction (90 d) [c]
Hogg et al. [9]	2016	Single center Retrospective Cohort	United States	NS	133	GS	Robotic Whipple	Postoperative Pancreatic Fistula	Nil
MacKenzie et al. [11]	2015	Multicenter Prospective Cohort	United Kingdom	20 (32%)	171	GS	Lap Right and left Hemicolectomy	Surgical Complications (30 d)	Nil
Birkmeyer et al.[7]	2013	Multicenter Retrospective Cohort	United States	15 (NS)	10242	GS	Lap Gastric Bypass	Complications (30 d)	Mortality (30 d) Readmission (30 d) Reoperation (30 d)

	ED visits (30 d)

[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

NS: Not Specified, GS: General Surgery, Lap: Laparoscopic; ED: Emergency department, EBWL%: Excess Body Weight Loss %

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.[13]	Modified OSATS [b]	Whole Procedure	Peer Raters	Yes	No	No	1-2 videos submitted by each surgeon
Brajcich et al.[71]	ASCRS[c] Video Assessment Tool	NA	Peer Raters & Experts	NA	NA	NA	1 video submitted by each surgeon
Stulberg et al. [72]	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. [12]	LapTMEpt Performance Assessment Tool	Whole Procedure	Expert	Yes	NA	Yes	1 video per patient
Fecso et al. [73]	OSATS GERT [d]	Critical parts of procedure	Experts	Yes	No	Yes	1 video per patient
Goldenberg et al. [10]	GEARS[e] GERT	Whole Procedure	Experts	Yes	No	Yes	1 video per patient
Scally et al. [74]	Modified OSATS	Critical parts of procedure	Peer Raters	Yes	Yes	No	1 video submitted by each surgeon
Paterson et al. [75]	VELP-Score [f]	Critical part of procedure	Experts	Yes	Yes	NA	1 video per patient
Hogg et al. [9]	Modified OSATS Technical Scoring Pancreaticojejunostomy	Critical parts of procedure	Experts	Yes	No	Yes	1 video per patient
MacKenzie et al.[11]	Competency Assessment Tool	Whole Procedure	Experts	Yes	No	Yes	1 video per patient
Birkmeyer et al. [7]	Modified OSATS	Critical parts of procedure	Peer Raters	Yes	Yes	No	1 video submitted by each surgeon

[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 Prostactomy Score
NA= Not Available

Table 3: Study Quality Assessment for Primary Outcomes

Author	Design	the Newcastle-Ottawa Scale [a]			
		Selection	Comparability	Outcome/Exposure	Risk of Bias [b]
Varaban et al.[13]	Multicenter Retrospective Cohort	☆☆	☆☆	☆☆	Low Risk of Bias
Brajcich et al.[71]	Multicenter Retrospective Cohort	☆☆	☆☆	☆☆	Low Risk of Bias
Stulberg et al. [72]	Multicenter Prospective Cohort	☆☆	☆☆	☆☆	Low Risk of Bias
Curtis et al. [12]	Multicenter Prospective Cohort	☆☆☆		☆☆	High Risk of Bias
Fecso et al. [73]	Multicenter Retrospective Cohort	☆☆	☆☆	☆	High Risk of Bias
Goldenberg et al. [10]	Single center Prospective Case-Control	☆☆☆	☆☆	☆☆☆	Low Risk of Bias
Scally et al. [74]	Multicenter Retrospective Cohort	☆☆☆	☆☆	☆☆	Low Risk of Bias
Paterson et al. [75]	Single center Prospective Cohort	☆☆☆	☆☆	☆	High Risk of Bias
Hogg et al. [9]	Single center Retrospective Cohort	☆☆☆	☆☆	☆	High Risk of Bias
MacKenzie et al.[11]	Multicenter Prospective Cohort	☆☆☆☆	☆☆	☆	High Risk of Bias
Birkmeyer et al. [7]	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

Table 4: Association of Intraoperative Performance with Postoperative Outcomes

Author	Quality Assessment	Operation Assessed	Outcomes Assessed (duration)	Effect
Varaban et al.[13]	Low risk of bias	Lap Sleeve Gastrectomy	Complications (30 d)	Rho: 0.21, p=0.30
			Readmission (30 d)	Rates: 1.9% vs. 2.9%, p=0.25
			Reoperation (30 d)	Rates: 0.2% vs. 0.9%, p<0.0001*
			ED visits (30 d)	Rates: 8.6% vs. 8.2 %, p=0.57
			EBWL % (1 year)	58.8% vs. 56.1%, p<0.03*
Brajcich et al.[71]	Low risk of bias	Lap Right Hemicolectomy	Survival (5-year)	HR: 0.31 [0.18, 0.54]*
Stulberg et al. [72]	Low risk of bias	Lap Hemicolectomy	Complications (30 d)	MD: 5.1% [0.4%, 9.8%], p=0.03*
			Mortality (30 d)	MD: 0.3% [–0.4%, 0.9%], p=0.59
			Readmission (30 d)	MD: 1.5% [−0.9%, 4.0%], p=0.27
			Reoperation (30 d)	MD: 2.5% [0.5%, 4.6%], p=0.02*
			Skill-related morbidity (30 d)	MD: 6.5% [−0.8%, 13.8%], p=0.08
			Skill-unrelated morbidity (30 d)	MD: 2.9% [−4.2%, 9.9%], p=0.55
Curtis et al. [12]	High risk of bias	Lap Total Mesorectal Excision	Complications (30 d)	Rates: 23.3% vs. 55.3%, p=0.03*
			Reoperation (30 d)	Rates: 3.3% vs. 6.3 %, p=0.6
			Readmission (30 d)	Rates: 0% vs. 33.3%, p=0.19
			Overall Survival (2-4 years)	Rates: 8.7% vs.96.6%, p=0.46
			Cancer Recurrence (2-4 years)	Rates: 70.0 % vs. 74.2, p=0.46
Fecso et al. [73]	High risk of bias	Lap Gastrectomy	Complications (30 d)	Rho: 0.401, p=0.001*
Goldenberg et al. [10]	Low risk of bias	Robotic Assisted Radical Prostectomy	Continence (3 month)	OR = 0.55 [0.33, 0.91]*
Scally et al. [74]	Low risk of bias	Lap Gastric Bypass	EBWL % (1 year)	67.2% vs. 68.5%; p =0.86
			Patient Satisfaction (1 year)	IR: 90.3 % vs. 87.1%; p =0.05*
Paterson et al. [75]	High risk of bias	Extraperitoneal Lap Prostectomy	Continence (3 month)	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
			Continence (12 months)	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. [9]	High risk of bias	Robotic Whipple	Postoperative Pancreatic Fistula	OR: 0.82 [0.70, 0.96]*
MacKenzie et al.[11]	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. [7]	Low risk of bias	Lap Gastric Bypass	Complications (30 d)	Rates: 5.2% vs. 14.5%, p<0.001*
			Mortality (30 d)	Rates: 0.05% vs. 0.26%, p=0.01*
			Readmission (30 d)	Rates: 2.7% vs. 6.3%, p<0.001*
			Reoperation (30 d)	Rates: 1.6 % vs. 3.4%, p=0.01*
			ED visits (30 d)	Rates: 3.8 vs.10.2%, p=0.004*

*Indicated statistical significance

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 %

SUPPLEMENTAL DIGITAL CONTENT

eMethods 1: Study Protocol

Hypothesis

We hypothesize that video-based assessments are a prevalent form of intra-operative technical performance assessment and patient outcomes.

Design Plan

Study type: Meta-Analysis - A systematic review of published studies.

Blinding: No blinding is involved in this study.

Study Design

Objective and Research Question

The main objective of this systematic review is to identify and summarize the existing literature on the association between video based assessments of intraoperative technical performance and patient outcomes.

Other secondary questions that will be explored are:

What VBAs have been used and for which procedures (or part of procedure)?

PICO Question

P: Surgeons in practice

I: Intraoperative performance

C: None

O: Postoperative outcome

Methods:

The search strategy was created in concert with a librarian and peer-reviewed by a second independent clinical librarian. Databases that will be searched include Medline, Embase, the Cochrane Library, and Web of Science.

Reporting:

The literature search, review, selection and reporting process will be conducted as per the guidelines set by the 'Preferred reporting items for systematic re-views and meta-analysis (PRISMA). This Protocol will be registered at Open Science Framework (osf.io). before commencement of literature review process.

Eligibility:

–Inclusion Criteria

1990- July 2020

All surgical specialties

Original full text articles

Studies evaluating intraoperative skills for surgeons in practice

Studies that use standardized assessment tools to evaluate intra-operative technical performance (excluding other competencies such as cognitive and interpersonal skills) from a recording

Studies that link intra-operative technical performance to patient outcomes

–Exclusion Criteria

Studies evaluating surgical trainees intraoperative skills

Studies that solely rely on surrogate measures of technical skills such as postoperative imaging or pathological specimens

Studies with qualitative evaluation (lack of use of standard assessment tool or framework) of intra-operative technical performance

Case Reports, comments, news & editorials

Non-human studies

Selection of abstracts and full-text articles:

Two reviewers (X and Y) will independently assess titles, abstracts and selected full-text of the articles obtained through the literature review. In case of any discrepancy between the included and excluded article a third independent reviewer (Z) will be used to assess the article in question for inclusion. The PRISMA Flow Diagram will be included to track the number of records identified, included/excluded, and reasons for exclusions.

Evaluation of the methodological quality of the included studies:

Each article included in the final selection will be evaluated for its methodological quality using the Newcastle-Ottawa Scale-Education (NOS-E). NOS-E is a validated system developed for assessment of quality of non-randomized trials based on three domains: the selection of the study groups; the comparability of the groups; and the ascertainment of the exposure/outcome of interest for case-control or cohort studies respectively.

Sampling Plan

Existing Data

Registration prior to creation of data

Data collection procedures

Data Extraction:

Similar to the earlier stages X and Y will work independently for information extraction from the selected articles. If available, the following data items will be extracted from all studies selected for inclusion. Items not available will be noted and reported as missing in the final report:

-Study Demographics

Authors

Year of publication

Study Design

Surgical Specialty

Operation assessed

Number of citations

Journal IF

-Performance Assessment

Time-point of assessment (whole procedure, specific challenging steps of the procedure)

Mode of assessment

Video Assessment

Number of video submissions per subject

Edited vs non-edited videos

Submission selection (i.e was the video selected by the operating surgeon or randomly chosen)

Assessment tools used

Qualification/numbers of the raters

Blinding status of the raters

Rater training

Measurement properties of the instrument

Number of patients assessed

Number of surgeons assessed

-Outcomes

Clinical outcomes

Patient reported outcome measures

Time-point of outcomes assessed

(immediate (intraoperative), short term (in-hospital) and long term outcomes (post-discharge))

Number of patients assessed

Study conclusion

Statistical analysis (statistical test used, Power and sample size)

Sample size

Not Applicable to Systematic Review

Analysis Plan

This systematic review will be reported using a narrative synthesis approach. If studies are sufficiently homogenous (with respect to design, population, intervention, and outcome measures), results will be pooled into a meta-analysis.

eMethods 2: Search Strategy

Medline [Ovid] <1946 to August 11, 2020>

Ovid MEDLINE(R) ALL <1946 to August 11, 2020>		
#	Searches	Results
1	exp Video Recording/	41092
2	video*.ti,kf.	35171
3	video*.ab. /freq=2	34685
4	1 or 2 or 3	79794
5	exp *Specialties, Surgical/mt, st	17359
6	exp *surgical procedures, operative/mt, st	589962
7	((surger* or surgeon* or surgical*) adj5 (skill* or proficien* or techni*)).tw,kf.	112248
8	clinical competence/	93456
9	5 or 6 or 7 or 8	766059
10	(assess* or measur* or quantif* or evaluat* or impact* or quantif* or rating* or rated or vary* or varied or variation* or rank*).tw,kf.	9316748
11	9 and 10	318329
12	outcome assessment, health care/ or treatment outcome/	1043275
13	outcome?.ti,kf.	376718
14	outcome?.ab. /freq=2	569436
15	postoperative*.tw,kf.	542867

16	12 or 13 or 14 or 15	1942492
17	11 and 16	151885
18	4 and 17	1938
19	limit 18 to yr="1990 -Current"	1934
20	exp animals/ not (exp animals/ and humans/)	4724721
21	19 not 20	1887
22	limit 21 to english language	1725
23	limit 22 to (comment or editorial or letter)	7
24	22 not 23	1718
25	remove duplicates from 24	1717

Embase [Ovid] <1974 to 2020 August 11>

Embase <1974 to 2020 August 11>		
#	Searches	Results
1	exp videorecording/	88591
2	video*.ti,kw.	47383
3	video*.ab. /freq=2	50778
4	or/1-3	129187
5	exp *surgical technique/	532226
6	((surger* or surgeon* or surgical*) adj5 (skill* or proficien* or techni*)).tw,kw.	151649
7	5 or 6	654203
8	4 and 7	16439
9	(assess* or measur* or quantif* or evaluat* or impact* or quantif* or rating* or rated or vary* or varied or variation* or rank*).tw,kw.	12271685
10	8 and 9	6766
11	outcome assessment/ or treatment outcome/	1365924
12	outcome?.ti,kw.	599287
13	outcome?.ab. /freq=2	888485
14	postoperative*.tw,kw.	715563
15	or/11-14	2684335
16	10 and 15	2886

17	limit 16 to yr="1990 -Current"	2878
18	limit 17 to (conference abstracts or embase)	2598
19	limit 18 to (editorial or letter or note)	6
20	18 not 19	2592
21	remove duplicates from 20	2587

Cochrane Library 2020/08/12

ID	Search	Hits
#1	(video*):ti,ab,kw	20803
#2	(((surger* or surgeon* or surgical*) near/5 (skill* or proficien* or techni*))):ti,ab,kw	13684
#3	#1 and #2	670
#4	((assess* or measur* or quantif* or evaluat* or impact* or quantif* or rating* or rated or vary* or varied or variation* or rank*)):ti,ab,kw	1008668
#5	#3 and #4	575
#6	(outcome* or postoperative*):ti,ab,kw	622759
#7	#5 and #6	296

Note: there were 8 Cochrane reviews, and 288 clinical trials. For the purpose of this review, the trials were not exported.

Web of Science
Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC

Timespan=All years

# 1	ts=(video*)	368,260
# 2	ts=(((surger* or surgeon* or surgical*) near/5 (skill* or proficien* or techni*)))	101,603
# 3	#1 and #2	4,133
# 4	ts=((assess* or measur* or quantif* or evaluat* or impact* or quantif* or rating* or rated or vary* or varied or variation* or rank*))	20,457,516
# 5	#3 and #4	2,520
# 6	ts=(outcome* or postoperative*)	2,555,338
# 7	#5 and #6	1,100
# 8	#7 not pmid=(1* or 2* or 3* or 4* or 5* or 6* or 7* or 8* or 9* or 0*)	85

Legends for Medline (Ovid), Embase (Ovid) & CINAHL (Ebsco) are available on our website:

http://www.muhclibraries.ca/Documents/Database_Legends.pdf

eTable1: Criteria for The Newcastle-Ottawa Scale Regarding Star Allocation to Assess Quality of Studies

Cohort Studies		
Criteria	**Acceptable (star awarded)**	**Unacceptable (star not awarded)**
Representativeness of the Exposed Cohort	– All videos included of consecutive cases from cohort series	–Surgeons/patients volunteered to participate and submit video(s)
Selection of the Non-Exposed Cohort	–Drawn from same source as exposed cohort	–Drawn from different source
Ascertainment of Exposure	–Blinded video assessment (i.e., the assessor did not know who was being assessed)	–Nonblinded video assessment or not reported
Demonstration that Outcome of Interest Was Not Present at Start of Study	–Statement of no history of disease or incident earns a star. –Surgical complication, efficacy measures or death where outcome could not have existed prior to study.	– Outcome assessment included medical complications without statement that it was not present preoperatively.
Comparability of Cohorts on the Basis of the Design or Analysis[a]	–Accounted for the most (≥ 4) important confounders (2 stars) –Accounted for some (1-3) important confounders (1 star)	–Did not account for confounders
Assessment of Outcome	–Blinded outcome assessment (included prospective blinded data collection or use of record linkage such as use of databases)	–Nonblinded outcome assessment (self-report, no description)
Was Follow-Up Long Enough for Outcomes to Occur	–For short term complications, mortality and resource utilization: 30 days –For cancer survival: 2 years –Weight loss: ≥ 1 year	–No
Adequacy of Follow-up [b]	–Loss to follow-up <20%	Loss to follow-up not reported or >20%
Case Control Studies		
Is the case definition adequate?	– Blinded outcome assessment (included blinded data collection or use of record linkage such as use of databases)	–Nonblinded assessment (self-report, no description)
Representativeness of the cases	–All or a random sample of consecutive cases	–Surgeons/patients volunteered to participate and/or submit video
Selection of Controls	–Drawn from same source the cases	–Drawn from same source the cases
Definition of Controls	–Explicit statement that controls have no history of this outcome.	–No mention of history of outcome
Comparability of cases and controls on the basis of the design or analysis [a]	–Accounted for the most (≥ 4) important confounders (2 stars) –Accounted for some (1-3) important confounders (1 star)	–Did not account for confounders
Ascertainment of exposure	–Blinded video assessment	–Nonblinded video assessment

Same method of ascertainment for cases and controls	–Yes	–No
Non-Response rate [b]	–Non-response rate <20%	–Non-response rate >20% or not reported

[a] Potential confounder for the most commonly assessed outcomes–post-operative complications and resource utilization–were defined *a priori* as patient related factors (BMI, age), comorbidities (reported in form of ASA, CCI or comorbidities themselves) and surgery related factors (emergency operations, type of surgery). Visser A, Geboers B, Gouma DJ, Goslings JC, Ubbink DT. Predictors of surgical complications: A systematic review. *Surgery.* 2015;158(1):58-65

[b] Dettori JR. Loss to follow-up. Evid Based Spine Care J. 2011;2(1):7-10.

eTable 2: Excluded Articles After Full Text Review

Reason for exclusion	Article
Not traced to full-text article	1. Zwart M, De Rooij T, Stommel M, Van den Boezem P, Wijsman J, Van der Schelling G, Schreinemakers J, Daams F, Zonderhuis B, Kazemier G, Mieog S. A nationwide training program for robotic pancreatoduodenectomy (LAELAPS-3): analysis of the first trained surgeons and first 87 patients. HPB. 2020 Jan 1;22:S320. 2. Kanters AE, Evilsizer S, Hendren S, Dimick JB, Byrn JC. Correlation of colorectal surgical skill with patient outcomes: a cautionary tale. Diseases Of The Colon & Rectum 2020 Jun 1 (Vol. 63, No. 6, pp. E92-E92). Two Commerce SQ, 2001 Market St, Philadelphia, PA 19103 USA: Lippincott Williams & Wilkins 3. Mingo S, Ma R, Nguyen J, Vanstrum E, Hung A. MP34-03 association of manual and automated performance metrics with urinary continence recovery after robot-assisted radical prostatectomy. The Journal of Urology. 2020 Apr;203(Supplement 4):e503-4. 4. Varban OA, Thumma JR, Telem DA, Obeid NR, Finks JF, Ghaferi AA, Dimick JB. Goldilocks Principle: Video Assessment of a Sleeve Gastrectomy That is "Just Right". Journal of the American College of Surgeons. 2019 Oct 1;229(4):e2-3. 5. Goldenberg M, Garbens A, Sadaat H, Finelli A, Singal R, Lee J, Grantcharov T. PD27-10 surgical performance as a predictor of functional and oncological outcomes in robotic prostatectomy. The Journal of Urology. 2019 Apr;201(Supplement 4):e484-5. 6. Beulens AJ, Brinkman WM, Meijer RP, Koldewijn EL, Van Basten JP, Vanmerrienboer JJ, Van Der Poel HG, Bangma CH, Wagner C. The use of multiple video assessment methods to determine the influence of surgical skill on potency and continency in patients after robot-assisted radical prostatectomy. European Urology Supplements. 2019 Sep 1;18(6):e2662. 7. Grober E, Goldenberg M, Elfassy M, Lorenzo A, Roberts M, Domes T, Mahdi M, AS Jewett M. PD58-02 validation of real-time, intra-operative, surgical competence (risc) assessments linked to clinically relevant patient outcomes: a model of competency assessment in urology. The Journal of Urology. 2018 Apr;199(4S):e1133-4. 8. Jung J, Dhir M, Zenati M, Novak S, Zureikat A, Zeh H, Hogg M. Predictors of delayed gastric emptying after robotic pancreatoduodenectomy: analysis of intraoperative techniques using video review. HPB. 2017 Apr 1;19:S10-1. 9. Goldenberg MG, Goldenberg SL, Grantcharov TP. Surgical technical performance impacts patient outcomes in robotic-assisted radical prostatectomy. J Urol. 2017;197 (4 Supplement 1):e699. 10. Ghani KR, Comstock B, Miller DC, Kim T, Linsell S, Lane BR, et al. Technical skill assessment of surgeons performing robot-assisted radical prostatectomy: Relationship between crowd sourced review and patient outcomes. J Urol. 2017;197 (4 Supplement 1):e609.

	11. Dunn RL, Peabody JO, Lane BR, Sarle R, Kim T, Brachulis A, et al. Music octave-composite measures to assess surgeon performance for robotic prostatectomy. J Urol. 2017;197 (4 Supplement 1):e1129-e30. 12. Paterson C, McLuckie S, Yew-Fung C, Anbarasan T, Tang B, Stolzenburg J, et al. Videotaping of surgical procedures and complications following extraperitoneal laparoscopic radical prostatectomy for clinically localised prostate cancer. J Endourol. 2016;30 (Supplement 2):A287-A8.
Not Linking VBAs of intraoperative technical performance to clinical outcome	13. Ploeg M, Keyzer-Dekker CM, Sloots CE, van de Ven CP, Meeussen C, Wijnen RM, Vlot J. Implementation of a quality control system for laparoscopic pyloromyotomy in hypertrophic pyloric stenosis: Hurdles and pitfalls. European Journal of Pediatric Surgery. 2019 Oct;29(05):443-8. 14. Ghodoussipour S, Reddy SS, Ma R, Huang D, Nguyen J, Hung AJ. An Objective Assessment of Performance during Robotic Partial Nephrectomy: Validation and Correlation of Automated Performance Metrics With Intraoperative Outcomes. The Journal of Urology. 2020 Dec 24:10-97. 15. Psaltis AJ, Li G, Vaezeafshar R, Cho KS, Hwang PH. Modification of the Lund-Kennedy endoscopic scoring system improves its reliability and correlation with patient-reported outcome measures. Laryngoscope. 2014;124(10):2216-23. 16. Mills JT, Hougen HY, Bitner D, Krupski TL, Schenkman NS. Does robotic surgical simulator performance correlate with surgical skill?. Journal of surgical education. 2017 Nov 1;74(6):1052-6. 17. Varban OA, Thumma JR, Carlin AM, Finks JF, Ghaferi AA, Dimick JB. Peer Assessment of operative videos with sleeve gastrectomy to determine optimal operative technique. Journal of the American College of Surgeons. 2020 Oct 1;231(4):470-7. 18. Han SU, Hur H, Lee HJ, Cho GS, Kim MC, Park YK, Kim W, Hyung WJ. Surgeon Quality Control and Standardization of D2 Lymphadenectomy for Gastric Cancer: A Prospective Multicenter Observational Study (KLASS-02-QC). Annals of Surgery. 2021 Feb 1;273(2):315-24.
VBA linked with comes in other procedures	19. Varban OA, Greenberg CC, Schram J, Ghaferi AA, Thumma JR, Carlin AM, Dimick JB, Michigan Bariatric Surgery Collaborative. Surgical skill in bariatric surgery: Does skill in one procedure predict outcomes for another?. Surgery. 2016 Nov 1;160(5):1172-81.
Lack of use of validated assessment tool	20. Arvidsson D, Berndsen FH, Larsson LG, Leijonmarck CE, Rimbäck G, Rudberg C, Smedberg S, Spangen L, Montgomery A. Randomized clinical trial comparing 5-year recurrence rate after laparoscopic versus Shouldice repair of primary inguinal hernia. British journal of surgery. 2005 Sep 1;92(9):1085-91.
Reviews	21. Prebay ZJ, Peabody JO, Miller DC, Ghani KR. Video review for measuring and improving skill in urological surgery. Nature Reviews Urology. 2019 Apr;16(4):261-7. 22. Shackelford S, Bowyer M. Modern metrics for evaluating surgical technical skills. Current Surgery Reports. 2017 Oct;5(10):1-0. 23. Varban OA, Ghaferi AA, Dimick JB. Using Video Analysis to Understand and Improve Technical Quality in Bariatric Surgery. Current Surgery Reports. 2017 Feb 1;5(2):6.

eTable 3: Risk of Bias Analysis Based on Newcastle-Ottawa Scale for Individual Studies

Cohort Studies	Outcomes	Representativeness of the Exposed Cohort	Selection of the Non-Exposed Cohort	Ascertainment of Exposure	Demonstration that Outcome of Interest Was Not Present at Start of Study	Comparability of Cohorts on the Basis of the Design or Analysis	Assessment of Outcome	Was Follow-Up Long Enough for Outcomes to Occur	Adequacy of Follow Up
Varaban et al.	Complications (30 d) Readmission (30 d) Reoperation (30 d) ED visits (30 d)	– Volunteered surgeons*	☆	☆	– No statement	☆☆ Adjusted for Comorbidities, Prior Hernia Repair, Age, Sex, Race, Insurance type, BMI	☆	☆	– Not disclosed
Varaban et al.	EBWL % (1 year)	– Volunteered surgeons*	☆	☆	☆	☆☆ Adjusted for Comorbidities, Prior Hernia Repair, Age, Sex, Race, Insurance type, BMI	☆	☆	– Not disclosed
Brajcich et al.	Survival (5-year)	– Volunteered surgeons*	☆	☆	☆	☆ Adjusted for Age, Sex, Race, CCI Stage, Operation Type	☆	☆	– Not disclosed
Stulberg et al.	Complications (30 d) Mortality (30 d) Readmission (30 d) Reoperation (30 d)	– Volunteered surgeons	☆	☆	– No statement	☆☆ Age, ASA, BMI, Sex, Race, Procedure type, Comorbidities, Preoperative laboratory values	☆	☆	– Not disclosed
Curtis et al.	Complications (30 d) Reoperation (30 d)	☆	☆	☆	– No statement	– No adjustments	☆	☆	– Not disclosed

	Readmission (30 d)								
Curtis et al.	Overall Survival (2-4 years) Cancer Recurrence (2-4 years)	☆	☆	☆	☆	– No adjustments	☆	☆	– Not disclosed
Fecso et al.	Complications (30 d)	– Volunteered surgeons	☆	☆	– No statement	☆☆ Adjusted for CCI, Gender, Type of Procedure	– No description of mode of outcome assessment	☆	– Not disclosed
Scally et al.	EBWL % (1 year)	– Volunteered surgeons	☆	☆	☆	☆☆ Adjusted for Age, Sex, Previous Thrombosis, Mobility status, Coronary Artery and Pulmonary Disease	☆	☆	– Not disclosed
Scally et al.	Patient Satisfaction (1 year)	– Volunteered surgeons	☆	☆	– No statement about baseline satisfaction given	☆ Adjusted for Age, Sex, Previous Thrombosis, Mobility status, Coronary Artery and Pulmonary Disease	– Outcomes are self-reported by patients	☆	– Not disclosed
Paterson et al.	Continence (3 mo) Continence (12 mo)	☆	☆	☆	– No statement	☆☆ Continence: Adjusted for Age, Grade, Stage, Size of prostate, CCI, History of Previous Prostate Surgeries,	– Continence (primary outcome) self-reported by patients through International Consultation on Incontinence	☆	– Loss to follow-up rare reported for database complication rate but not disclosed

						Socioeconomical status.	Questionnaire-Urinary Incontinence (ICQ-UI)		for outcomes reported by Patient Reported Outcome
Hogg et al.	Postoperative Pancreatic Fistula	– No description of nature of surgeon recruitment and only patients with available videos were included	☆	☆	☆	☆☆ Adjusted for BMI, Soft Gland, Duct Size, FRS/Braga Score (Pancreas Texture, Pancreatic Duct Diameter, Operative Blood Loss, And ASA Score)	☆	– Exact length of follow up not recorded	– Not disclosed
MacKenzie et al.	Surgical Complications (30 d)	☆	☆	☆	☆	☆☆ Adjusted for Age, BMI, ASA, Fitness grade, Tumour Stage, Previous Abdominal Surgery and Resection Type	– Outcomes are self-reported by surgeons	☆	– Not disclosed
Birkmeyer et al.	Complications (30 d) Mortality (30 d) Readmission (30 d) Reoperation (30 d) ED visits (30 d)	– Volunteered surgeons	☆	☆	– No statement	☆☆ Adjusted for Age, Sex, Previous Thrombosis, Mobility status, Coronary Artery and Pulmonary Disease	☆	☆	– Not disclosed
Case-Control		**Is the case definition**	**Representativeness of the cases**	**Selection of Controls**	**Definition of Controls**	**Comparability of cases and controls on the basis of**	**Ascertainment of exposure**	**Same method of ascertain**	**Non-Response rate**

Studies		**adequate?**				**the design or analysis**		**ment for cases and controls**	
Goldenberg et al.	Continence (3 mo) Continence (12 mo)	– Self-reported continence outcomes	☆	☆	☆	☆☆ Adjusted for Age, BMI, Prostate Weight	☆	☆	☆

Reasons for starts not awarded for each criterion is outlined

CHAPTER 4: Validity of intraoperative assessment tools for video-based self-assessment by general surgery trainees in laparoscopic cholecystectomy

4.1 Preamble to Manuscript 2

The systematic review outlined in Chapter 3 of this manuscript contributed evidence regarding the relationship between technical performance as measured through video-based assessment and surgical outcomes, supporting the association between greater intraoperative technical performance and lower perioperative complications and reoperations. Therefore, our findings suggest that intraoperative performance analysis using video-based assessment represents a promising approach to surgical quality-improvement, surgical training through feedback, benchmarking, and coaching.[6, 81]

These findings have further highlighted the importance of technical learning and proficiency during formal surgical training. Restriction of working hours during residency has reduced the operative exposure of surgical trainees, and almost one third of surgical graduates do not feel confident in their ability to perform certain procedures independently.[44] Hence, enhancement of training both inside and outside the operating room is crucial.[47] Review of video-recording of surgical procedures has offered new opportunities to trainees to extend technical learning to outside the operating room.[47] While video-assisted structured feedback by expert surgeons significantly improves laparoscopic skill acquisition in surgical trainees,[48] this method is resource intensive and has limited feasibility outside of research settings.[48]

Self-assessment is an integral part of lifelong medical experiential learning. However, systematic reviews report mixed results regarding the accuracy of trainee self-assessment.[51, 89, 90] These shortcomings can be remedied by using video recordings and implementing guided self-reflection strategies based on more robust intraoperative performance standards.[51, 55] Evidence supports the utility of guided self-reflection to improve performance in non-medical fields such as sports and music;[52] however, the value of video-based self-assessment in enhancing surgical skill acquisition should be investigated. This requires video-based assessment tools with validity evidence to allow reliable performance self-assessment.[33] Manuscript 2 will examine the validity of two commonly used intraoperative assessment tools for video-based self-assessment by general surgery trainees in laparoscopic cholecystectomy.

MANUSCRIPT 2: Validity of intraoperative assessment tools for video-based self-assessment by general surgery trainees in laparoscopic cholecystectomy

Validity of Video-Based Intra-Operative Self-Assessment by Surgical Trainees

Saba Balvardi, MD[1,2] Koorosh Semsar-Kazerooni, MS2[2] Pepa Kaneva, MSc[2] Carmen Mueller, MD Med[1,2] Melina Vassiliou, MD Med[1,2] Mohammed Al Mahroos, MD[1] Julio F Fiore Jr., PhD[1,2] Kevin Schwartzman, MD[3] Liane S Feldman, MD[1,2]

[1] Department of Surgery, McGill University, Montreal, Quebec, Canada

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

[3] Respiratory Division, Department of Medicine, McGill University and McGill International Tuberculosis Centre, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada

Corresponding author:

Liane S Feldman MD

Edward W. Archibald Professor and Chair, Department of Surgery, McGill University

1650 Cedar Ave, D6-156

Montreal, QC H3G 1A4

Tel: (514) 934-8044

Email: liane.feldman@mcgill.ca

Running head: Surgical trainee guided self-assessment

Funding: Fonds de la recherche en Sante du Quebec (FRSQ)

Word count: 2938

4.2 ABSTRACT

Introduction: Self-review of recorded surgical procedures offers new opportunities for trainees to extend technical learning outside the operating room. Valid tools for self-assessment are required prior to evaluating the effectiveness of video review in enhancing technical learning. Therefore, we aimed to contribute evidence regarding the validity of intraoperative performance assessment tools for video-based self-assessment by general surgery trainees when performing laparoscopic cholecystectomies.

Methods and Procedures: Using a web-based platform, general surgery trainees in a university-based residency program submitted video recordings of laparoscopic cholecystectomy procedures where they acted as the supervised primary surgeon. Operative performance was measured by the attending surgeon at the time of surgery using global and hybrid (global + procedure-specific) assessment tools (GOALS and OPRS, respectively) and entrustability level (O-SCORE). Trainees evaluated their own performance from video review using the same instruments. The validity of GOALS and OPRS for trainee self-assessment was investigated by testing the hypotheses that self-assessment scores correlate with (H1) expert assessment scores, (H2) O-SCORE, and (H3) procedure time and that (H4) self-assessment based on these instruments differentiates junior (postgraduate year (PGY)1-3) and senior trainees (PGY4-5), as well as (H5) simple (Visual Analogue Scale [VAS]≤ 4) versus complex cases (VAS>4). All hypotheses were based on previous literature, defined a priori, and were tested according to the COSMIN consensus on measurement properties.

Results: A total of 35 videos were submitted (45% female and 45% senior trainees) and self-assessed. Our data supported 2 out of 5 hypotheses (H1 and H4) for the GOALS tool and 3 out of 5 hypotheses (H1, H4 and H5) for the OPRS, for trainee self-assessment.

Conclusions: OPRS, a hybrid operative assessment tool, was better able to differentiate between groups expected to have different levels of intraoperative performance, compared to GOALS, a global assessment tool. Given the interest in video-based learning, there is a need to further develop valid procedure-specific tools to support video-based self-assessment by trainees in a range of procedures.

Key Words: Video-based assessment, Self-assessment, Validity, Intraoperative assessment tool

4.3 INTRODUCTION

Evidence suggests that surgical technique and skills directly influence safety and patient outcomes.[7, 11] A recent study has shown that almost one third of surgical graduates do not feel confident in their ability to perform certain procedures independently.[44] Modern challenges to surgical education includes restriction of working hours as well as the COVID-19 pandemic that reduced trainees' exposure to elective surgery through mandated cessation of non-essential procedures.[45, 46] Hence, extension of the technical learning outside the operating room has become crucial.[49] Video-based assessment (VBA) of recorded operative procedures provides a new opportunity to measure surgeon performance while minimizing barriers related to direct in-theater evaluations. While video-assisted structured feedback by expert surgeons significantly improves laparoscopic skill acquisition in surgical trainees,[48, 83, 91] this method is resource intensive and may have limited feasibility outside research settings. Accordingly, there is growing interest in the potential role of guided self-assessment of videorecorded surgical procedures to address this procedural training gap.[47]

Self-assessment is an integral part of lifelong medical experiential learning. Evidence supports the utility of guided self-assessment to improve performance in non-medical fields such as sports and music.[52] However, systematic reviews report mixed results regarding the accuracy of trainee self-assessment.[51, 89, 90] These shortcomings can be mitigated by using video recordings and implementing guided self-assessment strategies based on more robust intraoperative performance standards.[51, 55] Video-based tools with evidence supporting their valid use for self-assessment are required before the value of video-based self-assessment in enhancing surgical skill acquisition can be accurately investigated.[33] Thus the aim of this study was to contribute

evidence regarding the validity of intraoperative assessment tools when used for formative video-based self-assessment by general surgery trainees performing laparoscopic cholecystectomies.

4.4 MATERIALS AND METHODS

Participants and Settings

This study is a single-centered prospective cohort study that took place at the adult hospital sites of the McGill University Health Center. This was a sub-study of a recently completed randomized controlled trial with data collected from August 2020 to August 2021 (Effect of video-based guided self-reflection on intraoperative skills: a pilot randomized controlled trial; ClinicalTrials.gov Identifier: NCT04643314). This sub-study was approved by our institutional review board (MUHC Ethics Approval ID: 2020-6348). Inclusion criteria were: (1) Postgraduate year (PGY) 2-5 trainees, (2) rotating through a General Surgery Clinical Teaching Unit, (3) performing elective or emergency laparoscopic cholecystectomies (4) and performing more than 70% of the procedure. PGY1 trainees and procedures where there was significant (more than 30%) supervising surgeon take over were excluded.

Measures and Procedures

Demographic data from the trainees (i.e., age, gender, postgraduate year, handedness, and number of previous laparoscopic and laparoscopic cholecystectomy procedures) and the characteristics of the operative procedures (diagnosis, urgency, and procedure time) were collected. Junior trainees were defined as PGY 2-3, and senior trainees were PGY 4-5. Total duration of the operation was defined as the time from skin incision to skin closure. Time to

dissect the triangle of Calot was defined as the time from completion of adhesiolysis to clipping the first structure in the triangle of Calot. Duration of dissection of the gallbladder bed was defined as the time from division of the last structure in the hepatocystic triangle until detachment of the gallbladder from the gallbladder bed. In case of rescue techniques such as antegrade cholecystectomy or subtotal cholecystectomy, only the total procedure time was collected. The operative times were measured based on these definitions by one of the authors blinded to the operator and operative case characteristics.

Each resident was given a data storage device (USB key) to record elective and emergency laparoscopic cholecystectomy procedures in which they acted as the supervised primary surgeon for a significant portion of the operation (more than 70% of the whole procedure for senior trainees or more than 70% of triangle of Calot dissection and/or gallbladder wall dissection for junior trainees). These videos were uploaded to a secure web-accessible platform (TheatOR) and any identifying features were removed by the platform. In addition to storage and facilitation of access to surgical videos, TheatOR segments the procedure into steps to enable more targeted review of different parts of the operation (i.e., preparation, triangle of Calot dissection, division of cystic structures, gallbladder separation, gallbladder packaging, extraction) and evaluates whether the critical view of safety was obtained. Trainees met with a member of the study team not involved in clinical supervision to receive coaching on the nature and the use of the intraoperative assessment tools and undergo rater training including demonstration of sample videos for low and high scores for each assessment items. The trainees were then asked to practice using the scales with calibrating videos in the same session. Subsequently, trainees were asked to review their own operating room recordings and

to assess themselves, entering their self-assessment scores directly into the TheatOR platform as shown in Supplementary Material 1.

Operative performance was assessed using two measures selected from tools identified by a systematic review of performance assessment tools for laparoscopic cholecystectomy[35]: Global Operative Assessment of Laparoscopic Skills (GOALS)[32], a global rating scale of laparoscopic skills and Operative Performance Rating System (OPRS)[92], a hybrid (generic + procedure-specific) assessment tool. Both GOALS and OPRS have been supported by evidence of validity for use in direct intraoperative and video-based evaluation by attending surgeons.[17] In the present study, the attending surgeon submitted their assessments using GOALS and OPRS immediately after the procedure with maximum allotted time of 72 hours. They also completed a post-procedural questionnaire to delineate the degree of involvement of the trainee in the procedure (0%–100%), case difficulty (using a visual analogue scale (VAS: 1-10)) and overall trainee entrustability using the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE) Entrustability Scale.[93]

GOALS is a global intraoperative performance assessment tool consisting of 5 items, each scored using a 5-point Likert scale where '1' represents the lowest level of performance and '5' is considered ideal performance. The total possible score ranges between 5 and 25.[32] The items evaluate depth perception, bimanual dexterity, efficiency, tissue handling and autonomy (Supplementary Material 2).[32] There is evidence for the validity of GOALS in direct intraoperative and video-based evaluation by attending surgeons.[17] OPRS is a hybrid (global and procedure specific) 10-item intraoperative assessment tool.[20] A rating scale of 1-5 is used to evaluate each item with a rating of four or higher indicating technical proficiency and

operative independence.[17] The final score is the mean score of the 10 items (Supplementary Material 3).[20] OPRS has been recommended for use in the setting of direct observation, but it can also be used in assessment of recorded procedures.[17] The O-SCORE is a valid and reliable intraoperative assessment of operative competence using a 5-level scale. The expert clinician ranks the trainee's independence from 1= "I had to do it" to 5= "I did not need to be there" (Supplementary Material 4).[23] This scale is only designed for direct observation and is not suited for video assessment.[23] This scale is included in trainee competency assessment by the Royal College of Physicians and Surgeons of Canada's competency-based medical education framework.

Validity Assessment

The validity of GOALS and OPRS tool as formative video-based self-assessment tools by general surgery trainees in laparoscopic cholecystectomy was evaluated based on COSMIN best practice guidelines for examining psychometric properties.[94] Based on previous literature, we hypothesized *a priori* that if GOALS and OPRS are valid video-based self-assessment tools for general surgery trainees in laparoscopic cholecystectomy, trainee self-assessment scores will correlate with (1) expert assessment scores [95] (2) O-SCORE Entrustability Scale [93, 96] and (3) procedure time[97] and that (4) self-assessment based on these instruments can differentiate junior (postgraduate year (PGY)1-3) from senior trainees (PGY4-5)[98], as well as (5) simple (VAS≤ 4) versus complex cases (VAS>4).[93]

Statistical Analysis

A sample size of 29 submissions was expected to be sufficient to detect moderate correlations, i.e. r=0.5 (as defined by COSMIN best practice guidelines)[94] with an α= 0.05 and β= 0.80. Continuous variables were reported using mean and standard deviation or median and interquartile range, as appropriate. Categorical variables were reported using frequencies and percentages.

Guidelines recommend that hypotheses testing be based on the expected direction and magnitude of differences or correlations rather than on sample size-dependent statistics, such as p values.[94] Hypotheses 1-3 were tested using Pearson or Spearman's rank Correlation where appropriate. We expected a moderate positive correlation (coefficient 0.3 to 0.5) between attending surgeon and trainee self-assessments, and a moderate negative correlation (coefficient -0.3 to -0.5) between trainee self-assessment and procedure time. Hypotheses 4-5 were tested using multiple linear regression while adjusting for gender (for hypotheses 4 and 5)[99], case complexity (for hypothesis 4) and PGY level (for hypothesis 5). We hypothesized that the magnitude of difference between groups would be equal to or greater than the minimal important difference (MID) of 2 for GOALS[32] and 0.3 for OPRS[38].[100] These MIDs were estimated based on distribution based method with differences above one-half of the standard deviation considered clinically meaningful.[101] To reduce the risk of bias arising from missing data, we used random-forest-based imputation of missing data using the missForest R package.[102] Statistical analyses were performed using RStudio (version 1.2.1577; RStudio, Inc., Boston, MA, USA).

4.5 RESULTS

A total of 35 intraoperative recordings of laparoscopic cholecystectomy procedures were submitted by 11 trainees. Two trainees refused self-assessment citing time constraints. The

trainee median age was 30 years, 45% were female and 45% were senior trainees (≥PGY 4). Fifty-five percent of the submitted cases were done by trainees who had been the primary operating surgeon in more than 20 laparoscopic cholecystectomies and 89% had been the primary operating surgeon in more than 50 laparoscopic procedures (Table 1).

Out of the 35 submitted intraoperative recordings of laparoscopic cholecystectomies 11 (37%) were of patients with acute cholecystitis and 8 (23%) were of patients with biliary colic. Twelve cases out of 35 (34%) were done on an emergency basis and 17 (48%) were deemed complex (VAS>4) by the attending supervising surgeon (Table 2). In 9 (26%) of these cases, the supervising attending surgeon took over for less than 30% of the duration of the procedure. Median length of procedure was 85 min, 20.5 min for dissection of the triangle of Calot and 10.4 min for dissection of the gallbladder bed (Table 2).

Table 3 summarizes the intraoperative attending surgeon assessment and the trainee video-based self-assessment scores for GOALS and OPRS. Median length of time to completion of the intraoperative assessments by the attending surgeon was 0 days. However, trainee median length of time to video based self-assessment was 10 days. The attending surgeon's GOALS and OPRS total scores were higher than the trainee's self-assessments (22 vs 18 [p=0.001] and 5.4 vs 3.7 [p< 0.001]; respectively).

Trainees' GOALS video self-assessment scores correlated with staff surgeon GOALS assessment scores (correlation coefficient 0.47) and mean self-assessment scores differed between senior versus junior trainees (adjusted mean difference 3.53 [3.06, 3.78]). However, they did not correlate with entrustability (O-SCORE) or total procedure time. GOALS scores were also not significantly different between complex versus simple procedures (Table 4). Trainees' OPRS self-

assessment scores correlated with staff surgeon assessment (correlation coefficient 0.35), differed between senior versus junior trainees (adjusted mean difference 0.51 [0.17, 0.84]) and differed between complex versus simple procedures (adjusted mean difference 0.39 [0.03 to 0.74]). However, they did not correlate with entrustability scores or procedure time (Table 4). Hypothesis 3 was further investigated by testing the correlation of the self-assessment scores with the duration of the dissection of the triangle of Calot and the dissection of the gallbladder from the liver bed separately. Both GOALS and OPRS self-assessment scores correlated with the duration of dissection of the gallbladder bed but not the triangle of Calot dissection.

There was an 11% rate of missing attending surgeon intraoperative assessment (Supplementary Material 5). However, sensitivity analysis by testing these hypotheses after imputation of missing data yielded similar findings (Supplementary Material 6).

4.6 DISCUSSION

GOALS and OPRS are two commonly used global and hybrid (global + procedure specific) intraoperative assessment tools in laparoscopic cholecystectomy. There is evidence for their validity as formative assessment tools for surgical trainees evaluated by attending surgeons.[35] In this study we contribute evidence regarding the validity of their use for video-based self-assessment by general surgical trainees. Of the 5 *a priori* hypotheses tested for validity, 2 were supported for GOALS while 3 were supported for OPRS, suggesting stronger support for the use of self-assessment tools with procedure specific items in this context.

Trainees' GOALS self-assessment scores correlated with expert GOALS assessment scores[95] and self-assessment scores were significantly higher in senior surgical trainees (PGY 4-5) compared

to junior trainees (PGY 2-3)[98]. In contrast to previous literature demonstrating that intraoperative technical skill scores obtained by direct observation by expert surgeons correlate with entrustability score,[96] procedure duration,[97] and operative case complexity,[93] these were not observed in our study for GOALS self-assessment scores. Trainees' OPRS self-assessment scores correlated with expert OPRS assessment scores[95] and self-assessment scores were significantly higher in senior surgical trainees (PGY 4-5) compared to junior trainees (PGY 2-3)[98] and in simple (VAS≤ 4) compared to more complex cases (VAS>4).[93] However, the previously demonstrated correlation between intraoperative technical skill assessed by attending surgeons and entrustability score[96] and procedure duration[97] were not detected using OPRS self-assessment scores.

Neither OPRS nor GOALS self-assessment scores correlated with the O-SCORE evaluating entrustability, despite studies reporting correlation between expert assessment scores and O-SCORE.[96] This could be due to inherent differences between the constructs that these tools are designed to measure. O-SCORE is a tool that is designed to assess surgical competence (i.e., technical skills, cognitive skills and non-technical skills including communication and leadership) and hence readiness for independent performance of a procedure.[93] In contrast, the assessment items in OPRS and GOALS are largely directed towards technical skills performance with one or two elements assessing cognitive skills (namely elements evaluating flow of the operation or trainees' autonomy).[35] This is supported by previous studies showing that self-assessment of cognitive tasks to be fundamentally different and less accurate than that of more objective technical tasks in trainees.[51] Furthermore, O-SCORE assessment incorporates an established external reference criterion (independent performance of a procedure as an

attending surgeon) but OPRS and GOALS items are susceptible to relative scoring by trainees based on training level (i.e., I did well as a junior resident in an emergency case), especially in more junior trainees who lack the full range of surgical skills. This can in turn result in end-aversion bias (avoidance of low scores during self-assessment due to an incorrect external reference).[103] Therefore, the discrepancy between our findings and the literature (that reported correlation between O-SCORE and OPRS) can be due to the significantly lower risk of end-aversion bias and superior cognitive task assessment in expert attending assessors compared to trainee self-assessment in our study.

Similarly, neither OPRS nor GOALS self-assessment scores correlated strongly with total procedure time, unlike what was previously reported in the literature.[97] In our analysis, procedure length was defined *a priori* as the time from skin incision to skin closure. We performed a sensitivity analysis looking at the association of self-assessment score with duration of dissection of the triangle of Calot and duration of dissection of the gallbladder bed separately. We observed a significant inverse correlation between self-assessment scores and time for dissection of the gallbladder bed. We hypothesize that the lack of correlation with total operative duration can be due to the variations in operative characteristics such as difficulty obtaining intra-abdominal access, presence of intra-abdominal adhesions, and gallbladder extraction that are independent from technical skills but can affect the procedure duration. Furthermore, previous studies have also suggested a significant disagreement between surgeons regarding when the 'critical view of safety' is achieved or when dissection of the triangle of Calot can be deemed adequate.[104, 105] Therefore, the lack of correlation of the self-

assessment scores and the duration of dissection of the triangle of Calot can reflect the variability in defining the endpoint of this dissection between supervising attending surgeons.

A systematic review of self-assessment of technical tasks in surgery by Zevin *et al.* reported mixed results regarding the accuracy of trainee self-assessment.[51] These findings have been partly attributed to methodological limitations of previous studies and to factors such as recall bias (i.e. poor recall of intraoperative events by trainees after the fact).[51] Cognitive factors such as 'memory bias' have also been reported to affect accuracy of self-assessment. Memory bias is a defense mechanism that encourages poor recall of personal failures to decrease unhappiness and despair.[50] The use of intraoperative recording review and valid and reliable assessment tools with unambiguous behavioral anchors have been associated with improved accuracy of self-assessment.[51, 53] Furthermore, video-based self-reflection has been found to readily address factors such as recall bias and memory bias, and valid assessment tools with clear performance anchors have the potential to address the lack of accuracy and inconsistency in interpretation of items.[51, 55] Our findings corroborated these previously outlined observations as we observed that OPRS (as a hybrid assessment tool that includes procedure specific performance anchors) had stronger evidence of validity as a self-assessment tool compared to GOALS (a global assessment tool).

However, although GOALS and OPRS self-assessment scores significantly correlated with expert scores, they were consistently lower than expert scores. This discrepancy could be due to participant characteristics such as self-confidence, level of training or trainee gender.[51] Trainees who are women and trainees with low self-confidence have been reported to more frequently underestimate their performance.[51] Furthermore, rater training is an important avenue for

minimizing information bias as it improves accuracy and reliability of assessment using standardized tools.[87] In our study we used a personal session to familiarize the trainees with the assessment tools and performance anchors, and provide them with video examples. This method is formally known as 'Performance Dimension Training'.[87] Even though this method has been shown to improve rater accuracy, raters remain susceptible to the 'drift effect' where assessment accuracy can decline with time after initial training.[106] In future, providing longitudinal self-assessment feedback by comparison to expert assessments (i.e., Frame-of-reference Training) can lead to more significant and sustained positive impact on self-assessment accuracy.[107] Consequently, lack of adequate rater training or the drift effect could have introduced non-differential information bias in this study.[108]

The strength of our study lies in the robust methodology used for validity assessment. We followed COSMIN best practice guidelines and hypotheses were posed *a priori* to prevent reporting bias.[94, 109] We observed that the median time to completion of intraoperative attending assessment was 0 days, with 75% of evaluations being completed by 1 day after the procedure. decreasing the chance of recall bias of direct intraoperative assessments by attending surgeons. The median time to completion of trainee self-assessments was 10 days with a larger interquartile range. The use of intraoperative recording for self-assessment reduces concern about recall bias. However, since our data came from trainees who were interested in self-assessment, an element of selection bias cannot be excluded. Another limitation of our study is that the 35 videos analyzed were submitted by 11 trainees, introducing a clustering effect between submissions by the same trainee (i.e., values for videos obtained from the same trainee have a different relationship to one another than values for

videos obtained from different trainees). Lack of accounting for clustering of data through statistical methods can introduce type 1 error. [110] However, given the size of the clusters (with two to five videos submitted by a given trainee), cluster analysis is not recommended and hence it was not performed.[111] Hierarchical linear modeling (HLM)is another statistical solution to decreasing type 1 error when analysing clustered data. However, previous research has shown that this strategy in sparsely clustered data (cluster size < 5) is not recommended due to significant decrease in power.[112]

In summary, our study contributes evidence supporting the validity of GOALS and OPRS for formative trainee video-based self-assessment. There was stronger support for the use of OPRS, with 3 of 5 validity hypotheses supported, suggesting a potential advantage for assessments that include procedure-specific items compared to global assessments alone. These tools and their procedure-specific performance anchors can act as a guide for more accurate introspection and therefore may enhance their educational value for procedural learning. Given the reduced operative exposure of surgical trainees[44], use of these strategies to expand skills training outside the operating room is crucial.[47] Future research should focus on developing procedure-specific video-based assessment tools with robust measurement properties. This is an important step that will be required to investigate whether video self-review can improve procedural learning by surgical trainees.

DISCLOSURES

Saba Balvardi, MD: no conflict of interest

Koorosh Semsar-Kazerooni, MS: no conflict of interest

Pepa Kaneva, MSc: no conflict of interest

Carmen Mueller MD Med: no conflict of interest

Melina Vassiliou MD Med: no conflict of interest

Mohammed Al Mahroos, MD: no conflict of interest

Julio F Fiore Jr., PhD: no conflict of interest

Kevin Schwartzman, MD: no conflict of interest

Liane S Feldman, MD: no conflict of interest

4.7 REFERENCES

1. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. Oct 10 2013;369(15):1434-42. doi:10.1056/NEJMsa1300625
2. Mackenzie H, Ni M, Miskovic D, et al. Clinical validity of consultant technical skills assessment in the English National Training Programme for Laparoscopic Colorectal Surgery. *Br J Surg*. Jul 2015;102(8):991-7. doi:10.1002/bjs.9828
3. Friedell ML, VanderMeer TJ, Cheatham ML, et al. Perceptions of graduating general surgery chief residents: are they confident in their training? *J Am Coll Surg*. Apr 2014;218(4):695-703. doi:10.1016/j.jamcollsurg.2013.12.022
4. Purdy AC, de Virgilio C, Kaji AH, et al. Factors Associated With General Surgery Residents' Operative Experience During the COVID-19 Pandemic. *JAMA Surg*. Aug 1 2021;156(8):767-774. doi:10.1001/jamasurg.2021.1978
5. Association CM. Clearing the backlog: The cost to return wait times to pre-pandemic levels Deloitte LLP and affiliated entities. Accessed June 8, 2021, 2021. https://www.cma.ca/sites/default/files/pdf/Media-Releases/Deloitte-Clearing-the-Backlog.pdf
6. Hamad GG, Brown MT, Clavijo-Alvarez JA. Postoperative video debriefing reduces technical errors in laparoscopic surgery. *Am J Surg*. Jul 2007;194(1):110-4. doi:10.1016/j.amjsurg.2006.10.027
7. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP. Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Ann Surg*. Aug 2015;262(2):205-12. doi:10.1097/SLA.0000000000001214
8. Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc*. Dec 2007;21(12):2240-3. doi:10.1007/s00464-007-9356-z
9. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M. The impact of feedback of intraoperative technical performance in surgery: a systematic review. *BMJ Open*. Jun 15 2015;5(6):e006759. doi:10.1136/bmjopen-2014-006759
10. Green JL, Suresh V, Bittar P, Ledbetter L, Mithani SK, Allori A. The Utilization of Video Technology in Surgical Education: A Systematic Review. *J Surg Res*. Mar 2019;235:171-180. doi:10.1016/j.jss.2018.09.015
11. Liebermann DG, Katz L, Hughes MD, Bartlett RM, McClements J, Franks IM. Advances in the application of information technology to sport performance. *J Sports Sci*. Oct 2002;20(10):755-69. doi:10.1080/026404102320675611
12. Zevin B. Self versus external assessment for technical tasks in surgery: a narrative review. *J Grad Med Educ*. Dec 2012;4(4):417-24. doi:10.4300/JGME-D-11-00277.1
13. Stern J, Sharma S, Mendoza P, et al. Surgeon perception is not a good predictor of peri-operative outcomes in robot-assisted radical prostatectomy. *J Robot Surg*. Dec 2011;5(4):283-8. doi:10.1007/s11701-011-0293-4
14. Ganni S, Chmarra MK, Goossens RHM, Jakimowicz JJ. Self-assessment in laparoscopic surgical skills training: Is it reliable? *Surg Endosc*. Jun 2017;31(6):2451-2456. doi:10.1007/s00464-016-5246-6

15. Bull NB, Silverman CD, Bonrath EM. Targeted surgical coaching can improve operative self-assessment ability: A single-blinded nonrandomized trial. *Surgery*. Sep 27 2019;doi:10.1016/j.surg.2019.08.002
16. Ritter EM, Gardner AK, Dunkin BJ, Schultz L, Pryor AD, Feldman L. Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative performance assessment. *Surg Endosc*. Jul 2020;34(7):3176-3183. doi:10.1007/s00464-019-07089-y
17. Watanabe Y, Bilgic E, Lebedeva E, et al. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surg Endosc*. Mar 2016;30(3):832-44. doi:10.1007/s00464-015-4285-8
18. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. Jul 2005;190(1):107-13. doi:10.1016/j.amjsurg.2005.04.004
19. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. Oct 2005;138(4):640-7; discussion 647-9. doi:10.1016/j.surg.2005.07.017
20. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. Oct 2012;87(10):1401-7. doi:10.1097/ACM.0b013e3182677805
21. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. May 2012;21(4):651-7. doi:10.1007/s11136-011-9960-1
22. Bull NB, Silverman CD, Bonrath EM. Targeted surgical coaching can improve operative self-assessment ability: A single-blinded nonrandomized trial. *Surgery*. Feb 2020;167(2):308-313. doi:10.1016/j.surg.2019.08.002
23. Thanawala RM, Jesneck JL, Seymour NE. Education Management Platform Enables Delivery and Comparison of Multiple Evaluation Types. *J Surg Educ*. Nov - Dec 2019;76(6):e209-e216. doi:10.1016/j.jsurg.2019.08.017
24. Aggarwal R, Grantcharov T, Moorthy K, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg*. Jun 2007;245(6):992-9. doi:10.1097/01.sla.0000262780.17950.e5
25. Williams RG, Sanfey H, Chen XP, Dunnington GL. A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. *Ann Surg*. Jul 2012;256(1):177-87. doi:10.1097/SLA.0b013e31825b6de4
26. Ali A, Subhi Y, Ringsted C, Konge L. Gender differences in the acquisition of surgical skills: a systematic review. *Surg Endosc*. Nov 2015;29(11):3065-73. doi:10.1007/s00464-015-4092-2
27. Kim MJ, Williams RG, Boehler ML, Ketchum JK, Dunnington GL. Refining the evaluation of operating room performance. *J Surg Educ*. Nov-Dec 2009;66(6):352-6. doi:10.1016/j.jsurg.2009.09.005
28. Calland JF, Turrentine FE, Guerlain S, et al. The surgical safety checklist: lessons learned during implementation. *Am Surg*. Sep 2011;77(9):1131-7.

29. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. May 2003;41(5):582-92. doi:10.1097/01.mlr.0000062554.74615.4c
30. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC medical research methodology*. 2020;20(1):1-12.
31. Dl S. Norman gR. Health measurement scales. A practical guide to their development and use. Oxford University Press Oxford:; 2008.
32. Stefanidis D, Chintalapudi N, Anderson-Montoya B, Oommen B, Tobben D, Pimentel M. How often do surgeons obtain the critical view of safety during laparoscopic cholecystectomy? *Surg Endosc*. Jan 2017;31(1):142-146. doi:10.1007/s00464-016-4943-5
33. Sgaramella LI, Gurrado A, Pasculli A, et al. The critical view of safety during laparoscopic cholecystectomy: Strasberg Yes or No? An Italian Multicentre study. *Surg Endosc*. Jul 2021;35(7):3698-3708. doi:10.1007/s00464-020-07852-6
34. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med*. Oct 2005;80(10 Suppl):S46-54. doi:10.1097/00001888-200510001-00015
35. Ganni S, Botden S, Schaap DP, Verhoeven BH, Goossens RHM, Jakimowicz JJ. "Reflection-Before-Practice" Improves Self-Assessment and End-Performance in Laparoscopic Surgical Skills Training. *J Surg Educ*. Mar - Apr 2018;75(2):527-533. doi:10.1016/j.jsurg.2017.07.030
36. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. Fall 2012;32(4):279-86. doi:10.1002/chp.21156
37. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. Jan 2009;24(1):74-9. doi:10.1007/s11606-008-0842-3
38. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. *Journal of occupational and organizational psychology*. 1994;67(3):189-205.
39. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron Clinical Practice*. 2010;115(2):c94-c99.
40. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual*. 2018;78(1)
41. Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Med Res Methodol*. Mar 2 2005;5:10. doi:10.1186/1471-2288-5-10
42. Dalmaijer ES, Nord CL, Astle DE. Statistical power for cluster analysis. *arXiv preprint arXiv:200300381*. 2020;

Table 1: Characteristics of trainee operator

Variables	
Number of trainees	11
Age, years	30.0 (29.0, 31.5)
Gender	
Female	5 (45%)
Male	6 (55%)
Training level	
Junior trainees (PGY 2-3)	6 (55%)
Senior trainees (≥PGY 4)	5 (45%)
Handedness	
Right-handed	11 (100%)
Left-handed	0 (0%)
Previous laparoscopic cholecystectomy experience	
≤ 20	5 (45%)
> 20	6 (55%)
Previous laparoscopic experience	
≤ 50	2 (18%)
> 50	9 (89%)

Data is presented as median (IQR) or n (%)

IQR: Interquartile range; PGY: Postgraduate year

Table 2: Operative case characteristics

Variables	
Number of videos, n	35
Diagnosis, n (%)	
Acute Cholecystitis	13 (37%)
Biliary colic	8 (23%)
Chronic Cholecystitis	4 (11%)
Choledocholithiasis	3 (9%)
Gallbladder Polyp	3 (9%)
Pancreatitis	4 (11%)
Operative priority, n (%)	
Emergency	12 (34%)
Elective	23 (66%)
Complex procedure (VAS>4), n (%)	17 (48%)
Triangle of Calot dissection done by trainee, % (mean +/- SD)	89.3% $\pm$ 26.5%
Gallbladder bed dissection done by trainee, % (mean +/- SD)	96.9% $\pm$ 9.6%
Take-over by supervising surgeon, n (%)	
Yes	9 (26%)
No	26 (74%)
Procedure duration-minutes, median (IQR)	
Total procedure time	85.0 (66.0, 115.0)
Dissection of triangle of Calot duration	20.5 (16.1, 36.9)
Dissection of gallbladder bed	10.4 (7.8, 14.8)

IQR: Interquartile range; VAS: Visual Analogue Score; IQR: Interquartile Range

Table 3: Intraoperative expert assessment and video based self-assessment

Variables	**Intraoperative expert assessment Median (IQR)**	**Trainee self-assessment Median (IQR)**	**p value**
Time to assessment- days	0 (0-1)	10 (4-28)	NA
O-SCORE	4 (3,4)	NA	NA
GOALS	22 (19, 23)	18 (17,20)	0.001
Depth perception	5 (5, 5)	4 (4, 4)	< 0.001
Bimanual dexterity	4 (4, 5)	4 (3, 4)	0.01
Efficiency	4 (4, 5)	3 (3, 4)	< 0.001
Tissue handling	4 (4, 5)	4 (3, 4)	< 0.001
Autonomy	4 (3.2, 5)	4 (3, 4)	0.1
OPRS	4.5 (3.7, 4.9)	3.7 (3.3, 4)	< 0.001
Incision / Port Placement	5 (5, 5)	4 (4, 5)	0.001
Exposure	4 (4, 5)	4 (4, 4)	0.07
Cystic duct dissection	4 (4, 5)	4 (3, 4)	0.009
Cystic artery dissection	4 (4, 5)	4 (3, 4)	0.002
Gallbladder dissection	5 (4, 5)	4 (3, 4)	< 0.001
Instrument handling	4 (4, 5)	4 (3, 4)	0.003
Respect for tissue	5 (4, 5)	4 (3, 4)	< 0.001
Time and motion	4 (4, 5)	3 (3, 4)	< 0.001
Operation flow	4 (4, 5)	3 (3, 4)	< 0.001
Overall performance rating	5 (4, 5)	4 (3, 4)	< 0.001

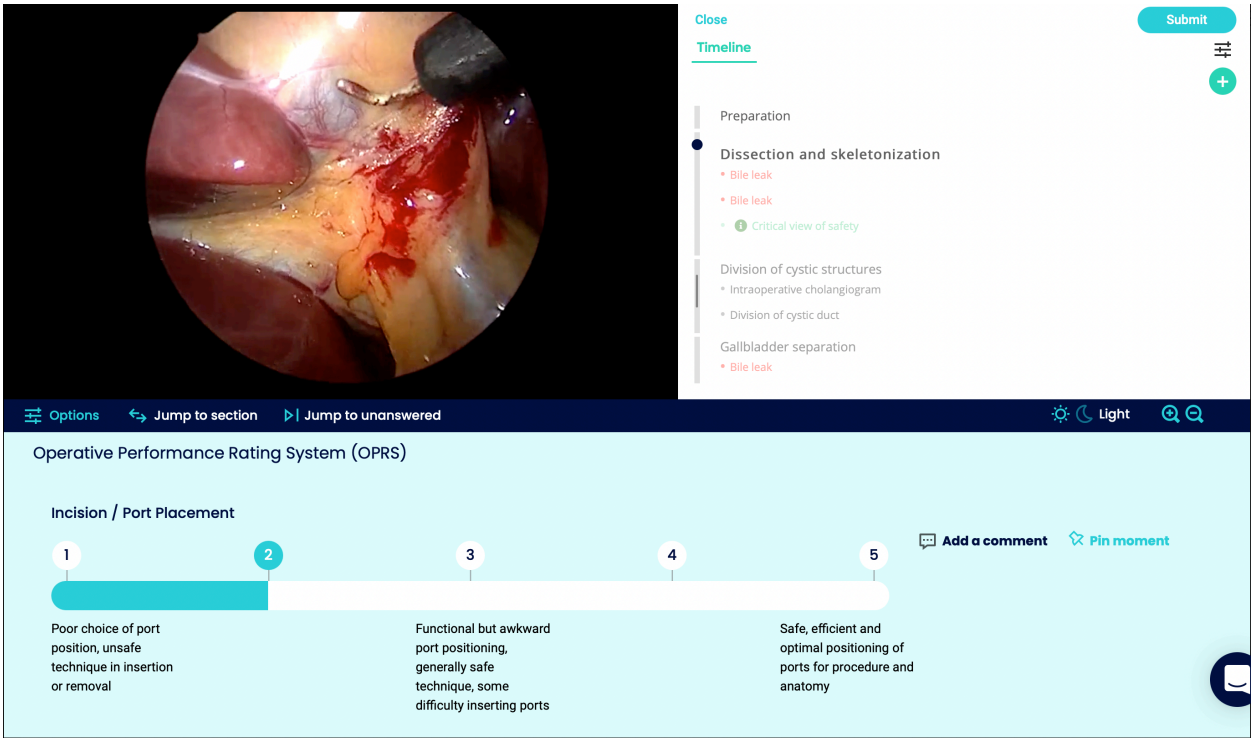
IQR: Interquartile Range; O-SCORE: Ottawa Surgical Competency Operating Room Evaluation; GOALS: Global Operative Assessment of Laparoscopic Skills; OPRS: Operative Performance Rating System; NA: Not Applicable

Table 4: Validity hypothesis testing

Hypothesis	GOALS				OPRS			
	Coefficient (95% CI) [a]	**Degrees of freedom**	**R^2**	**Hypothesis confirmed**	**Coefficient (95% CI)** [a]	**Degrees of freedom**	**R^2**	**Hypothesis confirmed**
(1) Correlation of self-assessment with expert score	**0.47**	33	NA	**Yes**	**0.35**	33	NA	**Yes**
(2) Correlation of self-assessment with expert entrustability score	0.17	33	NA	No	0.18	33	NA	No
(3) Correlation of self-assessment with total procedure time	–0.11	33	NA	No	–0.13	33	NA	No
a. Correlation with duration of TC dissection	–0.05	33	NA	No	–0.06	33	NA	No
b. Correlation with duration of GB bed dissection	**–0.41**	33	NA	**Yes**	**–0.32**	33	NA	**Yes**
(4) Mean difference in self-assessment score for senior vs. junior trainees	**3.53 (3.06, 3.78)**	28	0.33	**Yes**	**0.51 (0.17, 0.84)**	28	0.41	**Yes**
(5) Mean difference in self-assessment score for complex vs. simple cases	–1.56 (–3.40, 0.28)	28	0.33	No	**–0.39 (–0.74, –0.03)**	28	0.41	**Yes**

GOALS: Global Operative Assessment of Laparoscopic Skills; OPRS: Operative Performance Rating System; O-SCORE: Ottawa Surgical Competency Operating Room Evaluation; TC: Triangle of Calot, BG: gallbladder bed

[a] 95% CI is reported for regression coefficients



Supplementary Material 1: TheatOR Platform

Figure above demonstrates the trainee's view on the TheatOR platform during trainee's self-assessment of operative recording. The top left panel contains the intraoperative recording. The top right panel contains the automated segmentation of the procedure into steps to enable more targeted review of different parts of the operation. The bottom panel includes the intraoperative assessment tools that can be completed while viewing the intraoperative recording.

Supplementary Material 2: The Global Operative Assessment of Laparoscopic Skills (GOALS) assessment tool

Depth perception*

1. Constantly overshoots target, wide swings, slow to correct
2.
3. Some overshooting or missing of target, but quick to correct
4.
5. Accurately directs instruments in the correct plane to target

Bimanual dexterity*

1. Uses only one hand, ignores nondominant hand, poor coordination between hands
2.
3. Uses both hands, but does not optimize interaction between hands
4.
5. Expertly uses both hands in a complimentary manner to provide optimal exposure

Efficiency*

1. Uncertain, inefficient efforts; many tentative movements; constantly changing focus or persisting without progress
2.
3. Slow, but planned movements are reasonably organized
4.
5. Confident, efficient and safe conduct, maintains focus on task until it is better performed by way of an alternative approach

Tissue handling

1. Rough movements, tears tissue, injures adjacent structures, poor grasper control, grasper frequently slips
2.
3. Handles tissues reasonably well, minor trauma to adjacent tissue (ie, occasional unnecessary bleeding or slipping of the grasper)
4.
5. Handles tissues well, applies appropriate traction, negligible injury to adjacent structures

Autonomy

1. Unable to complete entire task, even with verbal guidance
2.
3. Able to complete task safely with moderate guidance
4.
5. Able to complete task independently without prompting

* The descriptors shown are the “anchor” descriptors for scores 1, 3, and 5.

Supplementary Material 3: The Operative Performance Rating System (OPRS) assessment tool

---

Procedure Specific Criteria

Incision / Port Placement

1. Poor choice of port position, unsafe technique in insertion or removal
2.
3. Functional but awkward port positioning, generally safe technique, some difficulty inserting ports
4.
5. Safe, efficient and optimal positioning of ports for procedure and anatomy

Exposure

1. Poor/inadequate pneumoperitoneum, camera angle and retraction with frequent loss of exposure
2.
3. Adequate establishment and maintenance of pneumoperitoneum, camera angle and retraction but with occasional loss of exposure and difficulty inserting ports
4.
5. Optimizes exposure of Calot's triangle, efficiently directs gallbladder retraction and camera to maintain exposure and pneumoperitoneum

Cystic Duct Dissection

1. Dissection of duct inadequate to place clips and divide safely
2.
3. Adequate but inefficient dissection, clips secure but spacing not ideal
4.
5. Expedient dissection, safe clip placement and duct division

Cystic Artery Dissection

1. Dissection of artery inadequate to place clips and divide safely, excessive hemorrhage, used more than 8 clips
2.
3. Adequate but inefficient dissection, clips secure but spacing not ideal.
4.
5. Expedient dissection, safe clip placement and artery division.

Gallbladder Dissection

1. Inefficient; did not cleanly remove gallbladder; excessive bile spillage; repeated
2.
3. Removed gallbladder intact but strayed from plane, somewhat inefficient, minimal bile spilled; extra cautery needed for liver bleeding
4.
5. Efficient; maintained clean plane between gallbladder and liver bed throughout, no parenchymal injury or bile spillage.

General Criteria

Instrument Handling

1. Tentative or awkward movements, *often* did not visualize tips of instrument or clips poorly placed
2.
3. Competent use of instruments, *occasionally* appeared awkward or did not visualize instrument tips
4.
5. Fluid movements with instruments *consistently* using appropriate force, keeping tips in view, and placing clips securely

Respect for Tissue

1. *Frequent* unnecessary tissue force or damage by inappropriate instrument use
2.
3. Careful tissue handling, *occasional* inadvertent damage
4.
5. *Consistently* handled tissue carefully (appropriately), minimal tissue damage

Time and Motion

1. Many unnecessary moves
2.
3. Efficient time and motion, some unnecessary moves
4.
5. Clear economy of motion, and maximum efficiency

Operation Flow

1. Frequent lack of forward progression; frequently stopped operating and seemed unsure of next move
2.
3. Some forward planning, reasonable procedure progression
4.
5. Obviously planned course of operation and anticipation of next step

Overall Performance (Rating of 4 or higher indicates technically proficient performance and that the resident is ready to perform operation independently)

1. Poor
2. Fair
3. Good
4. Very good
5. Excellent

Supplementary Material 4: O-SCORE Entrustability Scale

Level 1:

"I (supervising surgeon) had to do it"

i.e., requires complete hands-on guidance, did not do, or was not given the opportunity to do

Level 2

"I had to talk them through"

i.e., able to perform tasks but requires constant direction

Level 3

"I had to prompt them from time to time"

i.e., demonstrates some independence, but requires intermittent direction

Level 4

"I needed to be in the room just in case"

i.e., independence but unaware of risks and still requires supervision for safe practice

Level 5

"I did not need to be there"

i.e., complete independence, understands risks and performs safely, practice ready

Supplementary Material 5: Percent missing evaluations

Variables	Intraoperative Expert Assessment	Trainee Self-assessment
Enthrustability O-SCORE	4 (3,4)	NA
Missing	4 (11.4%)	
GOALS assessment	22 (19, 23)	18 (17,20)
Missing	4 (11.4%)	0 (0%)
OPRS assessment	4.5 (3.7, 4.9)	3.7 (3.3, 4)
Missing	4 (11.4%)	0 (0%)

Supplementary Material 6: Validity hypothesis testing sensitivity analysis with multiple imputation of missing data

Hypothesis	GOALS		OPRS	
	Coefficient (95% CI) [a]	**Hypothesis confirmed**	**Coefficient (95% CI)** [a]	**Hypothesis confirmed**
(1) correlation with expert assessment	**0.46**	**Yes**	**0.36**	**Yes**
(2) correlation with expert assessment O-SCORE	0.17	No	0.14	No
(3) correlation with procedure time	–0.10	No	–0.13	No
a. Correlation with duration of TC dissection	–0.01	No	–0.05	No
b. Correlation with duration of GB bed dissection	**–0.43**	**Yes**	**–0.33**	**Yes**
(4) Differentiating junior vs. senior residents	**3.59(1.84, 5.34)**	**Yes**	**0.44 (0.12, 0.75)**	**Yes**
(5) Differentiating simple vs. complex cases	–1.11 (–2.92, 0.71)	No	**–0.33 (–0.66, –0.001)**	**Yes**

GOALS: Global Operative Assessment of Laparoscopic Skills; OPRS: Operative Performance Rating System; O-SCORE: Ottawa Surgical Competency Operating Room Evaluation; TC: Triangle of calot, BG: gallbladder bed

[a] 95% CI is reported for regression coefficients

CHAPTER 5: Discussion

5.1 Summary of Findings

The first manuscript presented in this thesis aimed to synthesize knowledge about the association of intraoperative skills and patient outcomes, and to investigate the role of VBAs in measuring intraoperative performance. Therefore, we performed a systematic review of the available literature on the association between intraoperative technical performance and patient outcomes in practicing surgeons. 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.

Considering these findings, a potential role for VBA as a component of surgical training could be envisioned. While previous research supports the use of video-based review as part of one-on-one coaching, this is very resource intensive and likely not scalable. In other technical domains, such as music and sports, video self-review plays a role in deliberate practice. Ultimately, an intervention aimed to improve technical performance in trainees based on video-based self-assessment will require tools with robust measurement properties for that purpose. In the second manuscript we investigated the validity of previously available intraoperative

assessment tools for trainee formative self-assessment as a technical learning adjunct in laparoscopic cholecystectomy. Our study contributed evidence supporting the validity of GOALS and OPRS for formative trainee video-based self-assessment. There was stronger support for the use of OPRS, with 3 of 5 validity hypotheses supported, suggesting an advantage for assessments that include procedure specific items compared to global assessments alone. Our study further suggested that these tools and their procedure specific performance anchors can act as a guide for more accurate introspection and therefore may enhance their educational value for procedural learning. There findings are especially relevant given the recent decline in operative exposure of surgical trainees.[44] This makes the identification of effective strategies to expand skills training outside the operating room increasingly relevant.[47]

5.2 Discussion on Study Design and Systematic Bias

Both manuscripts include a discussion of the choice of study design and the limitations of these studies in term of random error and systematic (i.e., confounding, information, and selection) biases. This section will delve into more detail on these topics.

In the first manuscript, to thoroughly identify and summarize all available literature on the association between intraoperative technical performance measured using VBAs and patient outcomes, we used a systematic review (SR) study design. This study type follows a set of scientific methods that explicitly aim to identify, appraise, and synthesize all relevant literature within a set criteria (specified by inclusion and exclusion criteria of the review) to answer a particular question.[113] To improve the methodological rigour of this review and ascertain its reproducibility we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.[56] SRs are susceptible to systematic biases that are unique to this

study design including citation bias and publication bias. In our SR we used the expertise of a medical librarian to ensure an inclusive and comprehensive literature search to decrease the risk of citation bias.[114] However, given that 3 of the 31 studies that underwent full text review were identified through cross referencing (i.e., reviewing citations of included studies),[59] we cannot rule out citation bias which arises from more ready identification of studies with positive results that are more likely to be cited by other studies.[115]

We registered our systematic review protocol including the target outcomes and analysis plan *a priori* at Open Science Framework (osf.io/c29yb). Registering the protocol decreases the chance of selective reporting bias for reporting of outcomes and fishing for statistically significant results. Furthermore, clear *a priori* defined inclusion and exclusion criteria will decrease publication bias that arises from selective inclusion of studies that can fit a previously defined narrative.[116] Multiple publication bias is another form of publication bias where the same study or studies with significant overlap are published multiple times. This is usually more common in studies with positive outcomes.[114] Even though in our study we did not identify true duplicate publications, four out of 11 of the included studies were from the same center (Michigan Bariatric Surgery Collaborative) using the same outcomes database for two procedures (bariatric sleeve gastrectomy or gastric bypass).[7, 8, 13, 74] Finally, delay to publication or time lag bias is another form of publication bias that was identified in our study. Of the 23 studies that were eliminated during full text reviews, 12 were not included as they could not be traced to a full-text article. Due to the new interest and recent surge of publications in this field, publication bias especially in the form of delay to publication should be taken into consideration when interpreting the findings of our study. Moreover, we could not use a

statistical approach to more accurately quantify publication bias in our SR as meta-analysis was precluded due to heterogeneity.[117]

In our SR all the 13 included studies were observational studies (12 cohort studies and one case control study). The highest level of evidence for studies on exposure comes from randomized controlled trials (RCTs). Therefore, the preferred study design for our purpose would have been a SR and meta-analysis of RCTs. However, for our exposure of interest (operative skills of the operating surgeon) randomization of exposure would not be ethical. Therefore, a RCT will not be a feasible study design. Instead, a well-designed observational study will be the best possible study design in this field.

Our inclusion and exclusion criteria were the first step toward ensuring inclusion of high-quality studies in our SR. We included studies that used a performance assessment tool with published evidence supporting validity for their intended use and interpretation, to ensure a valid source of exposure assessment and decreased risk of measurement bias. Moreover, we only included studies with direct assessment of intraoperative skills through VBAs and excluded studies that relied solely on surrogate measures of technical performance such as postoperative imaging or examination of pathological specimens. This enabled us to reduce bias related to accuracy of exposure assessment, since surrogate measures can bring additional measurement bias. For example, a postoperative Xray of an open reduction and internal fixation of a tibial fracture can show perfect alignment and hardware placement but cannot demonstrate nerve injury.

Systematic bias usually arises from selection bias, information (or measurement) bias or confounding bias. In this section we will discuss sources of selection bias and measurement bias in our SR and confounding bias will be discussed in Section 5.3 “Discussion on Analysis”. In the

studies identified in our SR, several possible sources of measurement bias were identified that were related to variability in the features of VBAs. 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 published recommendations for minimizing measurement error when using VBAs.[8, 28] While most studies followed the recommendation to use blinded evaluators, rater qualification varied and was described as either "peer" or "expert" evaluation. Use of multiple peer raters (as opposed to 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.[23] The literature supporting peer VBA assessment in comparison to expert assessment (the default gold standard) has been mixed [29, 30] with supporting evidence for their use in evaluating simple tasks such as knot tying [86] and in the presence of added information such as intraoperative audio[30]. To our knowledge, no studies support the use of peer assessment using only visual feedback from VBA in complex procedures. Until more evidence is available for optimizing the accuracy of peer assessment, use of expert assessment should be prioritized in future studies and the accuracy of exposure (intraoperative technical skills) assessment in studies using peer assessment should be interpreted with this limitation in mind.

There was also a wide range of definitions for rater training in the identified studies, ranging from passive training based on descriptive manuals[12] to full training programs with continuous calibration of the assessors.[73] Only one of 5 studies that used peer assessors described any attempts at rater 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 towards 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.[8, 23, 87]

We also believe that selection bias could be another reason for the minor inconsistency in the association between intraoperative technical performance and outcomes between studies identified in our SR. Selection bias arises when individuals have different probabilities of being included in a study sample according to relevant study characteristics such as exposure or outcome. Eight of the eleven identified studies consistently reported participation rates below 35% of the invited surgeons. Another area of potential bias was the inclusion of patients based on the availability of intraoperative video submitted voluntarily by their surgeon versus a consecutive cohort of patients where video and outcome data were both available. These methods of patient and participant recruitment are not only susceptible to selection bias through volunteer effect, evaluating a surgeon based on a single video does not consider a surgeon's learning curve or the evolution of their technique throughout their years of practice.

Therefore, in the first manuscript we not only synthesized the available literature on the association between VBA of intraoperative skills and patient outcomes, but we also tried to delineate the shortcomings in study design of the studies reviewed and presented some recommendations for study design as a road map for future studies in this field.

The second manuscript presented in this thesis is a cross-sectional study aimed at contributing evidence regarding the validity of intraoperative assessment tools when used for formative video-based self-assessment by general surgery trainees performing laparoscopic cholecystectomies. Although cross-sectional studies are the study design of choice in assessing

psychometric properties of assessment tools, some elements in the study design can give rise to selection and measurement bias that will be discussed here. Of the 13 trainees who recorded their intraoperative performance in laparoscopic cholecystectomy, only 11 submitted their self-assessments. Furthermore, as this study was based on trainees who volunteered to perform the self-assessment, an element of selection bias cannot be ruled out.

Elements that can contribute to measurement bias in our study included recall bias (i.e. poor recall of intraoperative events by trainees after the fact).[51] Cognitive factors such as 'memory bias' have also been reported to affect accuracy of self-assessment. Memory bias is a defense mechanism that encourages poor recall of personal failures to decrease unhappiness and despair.[50] The use of intraoperative recording review and valid and reliable assessment tool with unambiguous behavioral anchors have been associated with improved accuracy of self-assessment.[51, 53] Furthermore, video-based self-reflection has been found to readily address factors such as recall bias and memory bias, and valid assessment tools with clear performance anchors have the potential to address the lack of accuracy and inconsistency in interpretation of items.[51, 55] Our findings corroborated these previously outlined observation as we observed that OPRS (as a hybrid assessment tool that includes procedure specific performance anchors) had stronger evidence supporting its validity as a self-assessment tool compared to GOALS (a global assessment tool).

Furthermore, rater training is an important avenue for minimizing measurement bias as it has been proven to improve accuracy and reliability of assessment using standardized tools.[87] In our study we used a one-on-one session to familiarize the trainees with the assessment tools and performance anchors, and provide them with video examples. This method is formally

known as 'Performance Dimension Training'.[87] Even though this method has been shown to improve rater accuracy, studies have shown that the raters are susceptible to the 'drift effect' where assessment accuracy can decline with time after initial training.[106] In the future, providing longitudinal self-assessment feedback by comparison to expert assessments (i.e., Frame-of-Reference Training) may lead to more significant and sustained positive impact on self-assessment accuracy.[107] Consequently, lack of adequate rater training or the drift effect could have introduced non-differential measurement bias to this study.[108]

5.3 Discussion of Analysis

Some elements of the statistical analysis presented in manuscripts outlined in this thesis will be further discussed in detail below.

The first manuscript presented in this study was a SR, and although this study was done with the intention to perform a meta-analysis, this was precluded due to study heterogeneity with respect to population, exposure, and outcome measures. Meta-analysis is a method of combining results across comparable studies to increase the power and decrease risk of type 2 error in the statistical analysis, however it is not recommended when there is significant heterogeneity between studies.[118] We used the narrative synthesis approach [68] to synthesize the findings from the included study. Furthermore, we used an established risk of bias assessment tool, the Newcastle-Ottawa Scale (NOS), for risk of bias assessment , to highlight the findings of each study in the context of its methodological limitations.[60] The nature of the measurement and selection bias present in the identified studies was previously discussed. In this section we will review the strategies used in the studies identified in our SR to minimize risk of confounding bias.

Confounding bias is one of the main risks of bias in observational studies. Confounding bias can be addressed in different stages of a study: through study design (used in all studies identified in this SR by restriction of case enrollment to a certain diagnosis or a certain procedure or matching used in one of the studies with case-control design)[10] or through post hoc analytic methods such as multivariate regression analysis. All multicenter studies included in this SR except one[12] used multivariate regression analysis to adjust for *a priori* identified confounders, however these confounders were limited to the available patient variables in the retrospective analyses. Furthermore, none of the studies reported their sample size calculations, therefore underpowered analysis cannot be ruled out. Furthermore, five of the included studies[7, 12, 13, 72, 74] tested multiple outcomes in absence of an *a priori* registered protocol. This can raise concerns about multiple testing.

The second manuscript was a validity study of the two intraoperative assessment tools (GOALS and OPRS) for trainee self-assessment for laparoscopic cholecystectomy based on a hypothesis testing approach. These hypotheses were based on previous literature and were set *a priori* to prevent reporting bias.[94] Furthermore, best practice guidelines recommend that hypothesis testing be based on the expected direction and magnitude of differences or correlations rather than on sample size-dependent statistics, such as p values.[94] Hypotheses 1-3 were tested using Pearson or Spearman's rank Correlation where appropriate. We expected a moderate positive correlation (coefficient 0.3 to 0.5) between the attending surgeon and self-assessment and a moderate negative correlation (coefficient -0.3 to -0.5) between self-assessment and procedure time. Hypotheses 4-5 were tested using multiple linear regression while adjusting for gender (for hypothesis 4 and 5)[99], case complexity (for hypothesis 4) and PGY level (for hypothesis 5).

We hypothesized that the magnitude of difference between groups would be equal to or greater than the minimal important difference (MID) of 2 for GOALS[32] and 0.3 for OPRS. [38, 100] These MIDs were estimated based on the well-established distribution based method with differences above one-half of the standard deviation considered clinically meaningful.[101]

We used multivariate linear regression to control for measured confounder variables such as gender and training level. Female trainees and senior trainees have been reported to more frequently underestimate their performance and more junior trainees are reported to overestimate their performace.[51] Other variables such as self-confidence has also been shown to significantly affect self-assessment accuracy but it was not measured in this study.[51] However, it should be mentioned that while hypotheses 4 and 5 were tested using multivariate linear regression where adjusting for confounders was a possibility, hypotheses 1-3 were tested using correlation coefficients and adjustments for these confounders were not statistically possible. Subgroup analysis (restriction of correlation analysis to subgroups such as females versus males or junior trainees versus senior trainees) as a method of investigating possible confounding effect of these variables was also precluded in our study given the small sample size and therefore limited power for subgroup analysis.

Furthermore, another limitation of our study is that the 35 videos analyzed were submitted by 11 trainees, introducing a clustering effect between submissions by the same trainee. Cluster analysis can be incorporated into regression models to account for clustered data as custering of data can introduce type 1 error when it is not considered in the analysis. [110] However, given the size of the clusters (with two to five videos submitted by a given trainee), cluster analysis is not recommended and hence it was not performed.[111] Hierarchical linear modeling (HLM) is

another statistical solution to decreasing type 1 error when analysing clustered data. However, previous research has shown that this strategy in sparsely clustered data (cluster size < 5) is not recommended due to significant decrease in power.[112]

The proportion of missing data was 10% in our cohort. Missing data can lead to selection bias if data is not missing at random and is associated with the exposure or the outcomes (i.e., the trainees with less accurate self-assessment were more likely to not fill out the self-assessment tool). Sensitivity analysis can be used to assess the risk of bias arising from missing data. We used random-forest-based imputation of missing data using the missForest R package [102] in RStudio for our sensitivity analysis and estimated the same results; this finding supported the robustness of our analysis.

CHAPTER 6: Conclusions and Future Directions

The first manuscript –a systematic review–contributed evidence regarding the relationship between technical performance as measured through video-based assessment and surgical outcomes, supporting the association between greater intraoperative technical performance and lower perioperative complications and reoperations. Long-term outcomes were less commonly investigated with mixed results. Furthermore, in this review we appraised the shortcomings of study design when using VBA (including but not limited to video selection, rater type and rater training) as a guide for future study design. We believe that future research should investigate the impact of technical performance and technical variation on postoperative outcomes in a more diverse range of procedures and examine the effectiveness of interventions to improve technical skill on patient outcomes.

In the second manuscript we contributed evidence supporting the validity of the GOALS and OPRS instruments for formative trainee video-based self-assessment. There was stronger support for the use of OPRS, suggesting an advantage for assessments that include procedure specific items compared to global assessments alone. These tools and their procedure specific performance anchors can act as a guide for more accurate introspection and therefore may enhance the educational value for procedural learning. Future research should focus on developing procedure-specific video-based assessment tools with robust measurement properties. This is an important step forward that will enable studies aiming to investigate whether video self-assessment can improve procedural learning by surgical trainees.

REFERENCES

1. Zegers M, de Bruijne MC, Wagner C, et al. Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. *Qual Saf Health Care*. Aug 2009;18(4):297-302. doi:10.1136/qshc.2007.025924
2. Kable AK, Gibberd RW, Spigelman AD. Adverse events in surgical patients in Australia. *Int J Qual Health Care*. Aug 2002;14(4):269-76. doi:10.1093/intqhc/14.4.269
3. Fabri PJ, Zayas-Castro JL. Human error, not communication and systems, underlies surgical complications. *Surgery*. Oct 2008;144(4):557-63; discussion 563-5. doi:10.1016/j.surg.2008.06.011
4. Dimick JB, Varban OA. Surgical video analysis: an emerging tool for improving surgeon performance. *BMJ Qual Saf*. Aug 2015;24(8):490-1. doi:10.1136/bmjqs-2015-004439
5. Feldman LS, Pryor AD, Gardner AK, et al. SAGES Video-Based Assessment (VBA) program: a vision for life-long learning for surgeons. *Surg Endosc*. Aug 2020;34(8):3285-3288. doi:10.1007/s00464-020-07628-y
6. Greenberg CC, Dombrowski J, Dimick JB. Video-Based Surgical Coaching: An Emerging Approach to Performance Improvement. *JAMA Surg*. Mar 2016;151(3):282-3. doi:10.1001/jamasurg.2015.4442
7. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. Oct 10 2013;369(15):1434-42. doi:10.1056/NEJMsa1300625
8. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The Effect of Technical Performance on Patient Outcomes in Surgery: A Systematic Review. *Ann Surg*. Mar 2017;265(3):492-501. doi:10.1097/SLA.0000000000001959
9. Hogg ME, Zenati M, Novak S, et al. Grading of Surgeon Technical Performance Predicts Postoperative Pancreatic Fistula for Pancreaticoduodenectomy Independent of Patient-related Variables. *Ann Surg*. Sep 2016;264(3):482-91. doi:10.1097/SLA.0000000000001862
10. Goldenberg MG, Goldenberg L, Grantcharov TP. Surgeon Performance Predicts Early Continence After Robot-Assisted Radical Prostatectomy. *J Endourol*. Sep 2017;31(9):858-863. doi:10.1089/end.2017.0284
11. Mackenzie H, Ni M, Miskovic D, et al. Clinical validity of consultant technical skills assessment in the English National Training Programme for Laparoscopic Colorectal Surgery. *Br J Surg*. Jul 2015;102(8):991-7. doi:10.1002/bjs.9828
12. Curtis NJ, Foster JD, Miskovic D, et al. Association of Surgical Skill Assessment With Clinical Outcomes in Cancer Surgery. *JAMA Surg*. Jul 1 2020;155(7):590-598. doi:10.1001/jamasurg.2020.1004
13. Varban OA, Thumma JR, Finks JF, Carlin AM, Ghaferi AA, Dimick JB. Evaluating the Effect of Surgical Skill on Outcomes for Laparoscopic Sleeve Gastrectomy: A Video-based Study. *Ann Surg*. Apr 1 2021;273(4):766-771. doi:10.1097/SLA.0000000000003385
14. Duclos A, Peix JL, Colin C, et al. Influence of experience on performance of individual surgeons in thyroid surgery: prospective cross sectional multicentre study. *BMJ*. Jan 10 2012;344:d8041. doi:10.1136/bmj.d8041
15. Biancari F, Mikkola R, Heikkinen J, Lahtinen J, Kettunen U, Juvonen T. Individual surgeon's impact on the risk of re-exploration for excessive bleeding after coronary artery

bypass surgery. *J Cardiothorac Vasc Anesth*. Aug 2012;26(4):550-6. doi:10.1053/j.jvca.2012.02.009
16. Johnston MJ, Singh P, Pucher PH, et al. Systematic review with meta-analysis of the impact of surgical fellowship training on patient outcomes. *Br J Surg*. Sep 2015;102(10):1156-66. doi:10.1002/bjs.9860
17. Nordback L, Parviainen M, Raty S, Kuivanen H, Sand J. Resection of the head of the pancreas in Finland: effects of hospital and surgeon on short-term and long-term results. *Scand J Gastroenterol*. Dec 2002;37(12):1454-60. doi:10.1080/003655202762671350
18. Gali BM, Madziga AG, Na'aya HU, Yawe T. Management of adult incisional hernias at the University of Maiduguri Teaching Hospital. *Niger J Clin Pract*. Sep 2007;10(3):184-7.
19. Koprowski P, Golec M, Dybek W, et al. Evaluation of nonunions of the tibia diaphysis - own experience. *Ortop Traumatol Rehabil*. May-Jun 2007;9(3):246-53.
20. Pascarella R, Cucca G, Maresca A, et al. Methods to avoid gamma nail complications. *Chir Organi Mov*. Apr 2008;91(3):133-9. doi:10.1007/s12306-007-0030-3
21. Vanderweele TJ. Surrogate measures and consistent surrogates. *Biometrics*. Sep 2013;69(3):561-9. doi:10.1111/biom.12071
22. Yanes AF, McElroy LM, Abecassis ZA, Holl J, Woods D, Ladner DP. Observation for assessment of clinician performance: a narrative review. *BMJ Qual Saf*. Jan 2016;25(1):46-55. doi:10.1136/bmjqs-2015-004171
23. Bilgic E, Valanci-Aroesty S, Fried GM. Video Assessment of Surgeons and Surgery. *Adv Surg*. Sep 2020;54:205-214. doi:10.1016/j.yasu.2020.03.002
24. Grenda TR, Pradarelli JC, Dimick JB. Using Surgical Video to Improve Technique and Skill. *Ann Surg*. Jul 2016;264(1):32-3. doi:10.1097/SLA.0000000000001592
25. Mackenzie H, Cuming T, Miskovic D, et al. Design, delivery, and validation of a trainer curriculum for the national laparoscopic colorectal training program in England. *Ann Surg*. Jan 2015;261(1):149-56. doi:10.1097/SLA.0000000000000437
26. Mori T, Kimura T, Kitajima M. Skill accreditation system for laparoscopic gastroenterologic surgeons in Japan. *Minim Invasive Ther Allied Technol*. 2010;19(1):18-23. doi:10.3109/13645700903492969
27. Pavlidis I, Zavlin D, Khatri AR, Wesley A, Panagopoulos G, Echo A. Absence of Stressful Conditions Accelerates Dexterous Skill Acquisition in Surgery. *Sci Rep*. Feb 11 2019;9(1):1747. doi:10.1038/s41598-019-38727-z
28. Scott DJ, Rege RV, Bergen PC, et al. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. *J Laparoendosc Adv Surg Tech A*. Aug 2000;10(4):183-90. doi:10.1089/109264200421559
29. Joosten M, Bokkerink GMJ, Verhoeven BH, Sutcliffe J, de Blaauw I, Botden S. Are Self-Assessment and Peer Assessment of Added Value in Training Complex Pediatric Surgical Skills? *Eur J Pediatr Surg*. Feb 2021;31(1):25-33. doi:10.1055/s-0040-1715438
30. Scully RE, Deal SB, Clark MJ, et al. Concordance Between Expert and Nonexpert Ratings of Condensed Video-Based Trainee Operative Performance Assessment. *J Surg Educ*. May - Jun 2020;77(3):627-634. doi:10.1016/j.jsurg.2019.12.016
31. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS. Crowd-Sourced Assessment of Technical Skill: A Valid Method for Discriminating Basic Robotic Surgery Skills. *J Endourol*. Nov 2015;29(11):1295-301. doi:10.1089/end.2015.0191

32. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. Jul 2005;190(1):107-13. doi:10.1016/j.amjsurg.2005.04.004
33. Ritter EM, Gardner AK, Dunkin BJ, Schultz L, Pryor AD, Feldman L. Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative performance assessment. *Surg Endosc*. Jul 2020;34(7):3176-3183. doi:10.1007/s00464-019-07089-y
34. Dougherty P, Kasten SJ, Reynolds RK, Prince ME, Lypson ML. Intraoperative assessment of residents. *J Grad Med Educ*. Jun 2013;5(2):333-4. doi:10.4300/JGME-D-13-00074.1
35. Watanabe Y, Bilgic E, Lebedeva E, et al. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surg Endosc*. Mar 2016;30(3):832-44. doi:10.1007/s00464-015-4285-8
36. Clark NV, Pepin KJ, Einarsson JI. Surgical skills assessment tools in gynecology. *Curr Opin Obstet Gynecol*. Oct 2018;30(5):331-336. doi:10.1097/GCO.0000000000000477
37. Hwang H, Lim J, Kinnaird C, et al. Correlating motor performance with surgical error in laparoscopic cholecystectomy. *Surg Endosc*. Apr 2006;20(4):651-5. doi:10.1007/s00464-005-0370-8
38. Kim MJ, Williams RG, Boehler ML, Ketchum JK, Dunnington GL. Refining the evaluation of operating room performance. *J Surg Educ*. Nov-Dec 2009;66(6):352-6. doi:10.1016/j.jsurg.2009.09.005
39. Educativa CCdEplE, Educational CtDSf, Testing P, et al. *Standards for Educational and Psychological Testing*. Amer Educational Research Assn; 1999.
40. Madani A, Vassiliou MC, Watanabe Y, et al. What Are the Principles That Guide Behaviors in the Operating Room?: Creating a Framework to Define and Measure Performance. *Ann Surg*. Feb 2017;265(2):255-267. doi:10.1097/SLA.0000000000001962
41. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. Feb 1997;84(2):273-8. doi:10.1046/j.1365-2168.1997.02502.x
42. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg*. Feb 2008;247(2):372-9. doi:10.1097/SLA.0b013e318160b371
43. Bilgic E, Al Mahroos M, Landry T, Fried GM, Vassiliou MC, Feldman LS. Assessment of surgical performance of laparoscopic benign hiatal surgery: a systematic review. *Surg Endosc*. Nov 2019;33(11):3798-3805. doi:10.1007/s00464-019-06662-9
44. Friedell ML, VanderMeer TJ, Cheatham ML, et al. Perceptions of graduating general surgery chief residents: are they confident in their training? *J Am Coll Surg*. Apr 2014;218(4):695-703. doi:10.1016/j.jamcollsurg.2013.12.022
45. Purdy AC, de Virgilio C, Kaji AH, et al. Factors Associated With General Surgery Residents' Operative Experience During the COVID-19 Pandemic. *JAMA Surg*. Aug 1 2021;156(8):767-774. doi:10.1001/jamasurg.2021.1978
46. Association CM. Clearing the backlog: The cost to return wait times to pre-pandemic levels Deloitte LLP and affiliated entities. Accessed June 8, 2021, 2021. https://www.cma.ca/sites/default/files/pdf/Media-Releases/Deloitte-Clearing-the-Backlog.pdf

47. Green JL, Suresh V, Bittar P, Ledbetter L, Mithani SK, Allori A. The Utilization of Video Technology in Surgical Education: A Systematic Review. *J Surg Res*. Mar 2019;235:171-180. doi:10.1016/j.jss.2018.09.015
48. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP. Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Ann Surg*. Aug 2015;262(2):205-12. doi:10.1097/SLA.0000000000001214
49. Hamad GG, Brown MT, Clavijo-Alvarez JA. Postoperative video debriefing reduces technical errors in laparoscopic surgery. *Am J Surg*. Jul 2007;194(1):110-4. doi:10.1016/j.amjsurg.2006.10.027
50. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med*. Oct 2005;80(10 Suppl):S46-54. doi:10.1097/00001888-200510001-00015
51. Zevin B. Self versus external assessment for technical tasks in surgery: a narrative review. *J Grad Med Educ*. Dec 2012;4(4):417-24. doi:10.4300/JGME-D-11-00277.1
52. Liebermann DG, Katz L, Hughes MD, Bartlett RM, McClements J, Franks IM. Advances in the application of information technology to sport performance. *J Sports Sci*. Oct 2002;20(10):755-69. doi:10.1080/026404102320675611
53. Ganni S, Botden S, Schaap DP, Verhoeven BH, Goossens RHM, Jakimowicz JJ. "Reflection-Before-Practice" Improves Self-Assessment and End-Performance in Laparoscopic Surgical Skills Training. *J Surg Educ*. Mar - Apr 2018;75(2):527-533. doi:10.1016/j.jsurg.2017.07.030
54. Blanch-Hartigan D. Medical students' self-assessment of performance: results from three meta-analyses. *Patient Educ Couns*. Jul 2011;84(1):3-9. doi:10.1016/j.pec.2010.06.037
55. Bull NB, Silverman CD, Bonrath EM. Targeted surgical coaching can improve operative self-assessment ability: A single-blinded nonrandomized trial. *Surgery*. Sep 27 2019;doi:10.1016/j.surg.2019.08.002
56. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. Mar 29 2021;372:n71. doi:10.1136/bmj.n71
57. Kelley WE, Jr. The evolution of laparoscopy and the revolution in surgery in the decade of the 1990s. *JSLS*. Oct-Dec 2008;12(4):351-7.
58. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. *J Clin Epidemiol*. Jul 2016;75:40-6. doi:10.1016/j.jclinepi.2016.01.021
59. Horsley T, Dingwall O, Sampson M. Checking reference lists to find additional studies for systematic reviews. *Cochrane Database Syst Rev*. Aug 10 2011;(8):MR000026. doi:10.1002/14651858.MR000026.pub2
60. Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. *Ottawa: Ottawa Hospital Research Institute*. 2011;
61. Wells G, Shea B, O'connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Ottawa: Ottawa Hospital Research Institute. oxford. asp; 2011.

62. Viale L, Allotey J, Cheong-See F, et al. Epilepsy in pregnancy and reproductive outcomes: a systematic review and meta-analysis. *Lancet*. Nov 7 2015;386(10006):1845-52. doi:10.1016/S0140-6736(15)00045-8
63. Sobhy S, Zamora J, Dharmarajah K, et al. Anaesthesia-related maternal mortality in low-income and middle-income countries: a systematic review and meta-analysis. *Lancet Glob Health*. May 2016;4(5):e320-7. doi:10.1016/S2214-109X(16)30003-1
64. Papola D, Ostuzzi G, Thabane L, Guyatt G, Barbui C. Antipsychotic drug exposure and risk of fracture: a systematic review and meta-analysis of observational studies. *Int Clin Psychopharmacol*. Jul 2018;33(4):181-196. doi:10.1097/YIC.0000000000000221
65. Wang B, An X, Shi X, Zhang JA. MANAGEMENT OF ENDOCRINE DISEASE: Suicide risk in patients with diabetes: a systematic review and meta-analysis. *Eur J Endocrinol*. Oct 2017;177(4):R169-R181. doi:10.1530/EJE-16-0952
66. Visser A, Geboers B, Gouma DJ, Goslings JC, Ubbink DT. Predictors of surgical complications: A systematic review. *Surgery*. Jul 2015;158(1):58-65. doi:10.1016/j.surg.2015.01.012
67. Dettori JR. Loss to follow-up. *Evid Based Spine Care J*. Feb 2011;2(1):7-10. doi:10.1055/s-0030-1267080
68. Popay J, Roberts H, Sowden A, et al. Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version*. 2006;1:b92.
69. Arvidsson D, Berndsen FH, Larsson LG, et al. Randomized clinical trial comparing 5-year recurrence rate after laparoscopic versus Shouldice repair of primary inguinal hernia. *Br J Surg*. Sep 2005;92(9):1085-91. doi:10.1002/bjs.5137
70. Mills JT, Hougen HY, Bitner D, Krupski TL, Schenkman NS. Does Robotic Surgical Simulator Performance Correlate With Surgical Skill? *J Surg Educ*. Nov - Dec 2017;74(6):1052-1056. doi:10.1016/j.jsurg.2017.05.011
71. Brajcich BC, Stulberg JJ, Palis BE, et al. Association Between Surgical Technical Skill and Long-term Survival for Colon Cancer. *JAMA Oncol*. Jan 1 2021;7(1):127-129. doi:10.1001/jamaoncol.2020.5462
72. Stulberg JJ, Huang R, Kreutzer L, et al. Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surgery*. 2020;19:19.
73. Fecso AB, Bhatti JA, Stotland PK, Quereshy FA, Grantcharov TP. Technical Performance as a Predictor of Clinical Outcomes in Laparoscopic Gastric Cancer Surgery. *Ann Surg*. Jul 2019;270(1):115-120. doi:10.1097/SLA.0000000000002741
74. Scally CP, Varban OA, Carlin AM, Birkmeyer JD, Dimick JB, Michigan Bariatric Surgery C. Video Ratings of Surgical Skill and Late Outcomes of Bariatric Surgery. *JAMA Surg*. Jun 15 2016;151(6):e160428. doi:10.1001/jamasurg.2016.0428
75. Paterson C, McLuckie S, Yew-Fung C, Tang B, Lang S, Nabi G. Videotaping of surgical procedures and outcomes following extraperitoneal laparoscopic radical prostatectomy for clinically localized prostate cancer. *J Surg Oncol*. Dec 2016;114(8):1016-1023. doi:10.1002/jso.24484
76. Varban OA, Sheetz KH, Cassidy RB, et al. Evaluating the effect of operative technique on leaks after laparoscopic sleeve gastrectomy: a case-control study. *Surg Obes Relat Dis*. Apr 2017;13(4):560-567. doi:10.1016/j.soard.2016.11.027

77. Ponce J. 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*. Mar 2018;267(3):e52. doi:10.1097/SLA.0000000000002076
78. Le AT, Huang B, Hnoosh D, et al. Effect of complications on oncologic outcomes after pancreaticoduodenectomy for pancreatic cancer. *J Surg Res*. Jun 15 2017;214:1-8. doi:10.1016/j.jss.2017.02.036
79. Park EJ, Baik SH, Kang J, et al. The Impact of Postoperative Complications on Long-term Oncologic Outcomes After Laparoscopic Low Anterior Resection for Rectal Cancer. *Medicine (Baltimore)*. Apr 2016;95(14):e3271. doi:10.1097/MD.0000000000003271
80. Beecher SM, O'Leary DP, McLaughlin R, Kerin MJ. The Impact of Surgical Complications on Cancer Recurrence Rates: A Literature Review. *Oncol Res Treat*. 2018;41(7-8):478-482. doi:10.1159/000487510
81. Greenberg CC, Ghousseini HN, Pavuluri Quamme SR, Beasley HL, Wiegmann DA. Surgical coaching for individual performance improvement. *Ann Surg*. Jan 2015;261(1):32-4. doi:10.1097/SLA.0000000000000776
82. 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
83. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M. The impact of feedback of intraoperative technical performance in surgery: a systematic review. *BMJ Open*. Jun 15 2015;5(6):e006759. doi:10.1136/bmjopen-2014-006759
84. Tam V, Zeh HJ, 3rd, Hogg ME. Incorporating Metrics of Surgical Proficiency Into Credentialing and Privileging Pathways. *JAMA Surg*. May 1 2017;152(5):494-495. doi:10.1001/jamasurg.2017.0025
85. Deijen CL, Velthuis S, Tsai A, et al. COLOR III: a multicentre randomised clinical trial comparing transanal TME versus laparoscopic TME for mid and low rectal cancer. *Surg Endosc*. Aug 2016;30(8):3210-5. doi:10.1007/s00464-015-4615-x
86. Deal SB, Lendvay TS, Haque MI, et al. Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. *Am J Surg*. Feb 2016;211(2):398-404. doi:10.1016/j.amjsurg.2015.09.005
87. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. Fall 2012;32(4):279-86. doi:10.1002/chp.21156
88. Bilgic E, Watanabe Y, McKendy K, et al. Reliable assessment of operative performance. *Am J Surg*. Feb 2016;211(2):426-30. doi:10.1016/j.amjsurg.2015.10.008
89. Stern J, Sharma S, Mendoza P, et al. Surgeon perception is not a good predictor of peri-operative outcomes in robot-assisted radical prostatectomy. *J Robot Surg*. Dec 2011;5(4):283-8. doi:10.1007/s11701-011-0293-4
90. Ganni S, Chmarra MK, Goossens RHM, Jakimowicz JJ. Self-assessment in laparoscopic surgical skills training: Is it reliable? *Surg Endosc*. Jun 2017;31(6):2451-2456. doi:10.1007/s00464-016-5246-6

91. Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc*. Dec 2007;21(12):2240-3. doi:10.1007/s00464-007-9356-z
92. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. Oct 2005;138(4):640-7; discussion 647-9. doi:10.1016/j.surg.2005.07.017
93. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. Oct 2012;87(10):1401-7. doi:10.1097/ACM.0b013e3182677805
94. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. May 2012;21(4):651-7. doi:10.1007/s11136-011-9960-1
95. Bull NB, Silverman CD, Bonrath EM. Targeted surgical coaching can improve operative self-assessment ability: A single-blinded nonrandomized trial. *Surgery*. Feb 2020;167(2):308-313. doi:10.1016/j.surg.2019.08.002
96. Thanawala RM, Jesneck JL, Seymour NE. Education Management Platform Enables Delivery and Comparison of Multiple Evaluation Types. *J Surg Educ*. Nov - Dec 2019;76(6):e209-e216. doi:10.1016/j.jsurg.2019.08.017
97. Aggarwal R, Grantcharov T, Moorthy K, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg*. Jun 2007;245(6):992-9. doi:10.1097/01.sla.0000262780.17950.e5
98. Williams RG, Sanfey H, Chen XP, Dunnington GL. A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. *Ann Surg*. Jul 2012;256(1):177-87. doi:10.1097/SLA.0b013e31825b6de4
99. Ali A, Subhi Y, Ringsted C, Konge L. Gender differences in the acquisition of surgical skills: a systematic review. *Surg Endosc*. Nov 2015;29(11):3065-73. doi:10.1007/s00464-015-4092-2
100. Calland JF, Turrentine FE, Guerlain S, et al. The surgical safety checklist: lessons learned during implementation. *Am Surg*. Sep 2011;77(9):1131-7.
101. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. May 2003;41(5):582-92. doi:10.1097/01.mlr.0000062554.74615.4c
102. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. *BMC medical research methodology*. 2020;20(1):1-12.
103. Dl S. Norman gR. Health measurement scales. A practical guide to their development and use. Oxford University Press Oxford:; 2008.
104. Stefanidis D, Chintalapudi N, Anderson-Montoya B, Oommen B, Tobben D, Pimentel M. How often do surgeons obtain the critical view of safety during laparoscopic cholecystectomy? *Surg Endosc*. Jan 2017;31(1):142-146. doi:10.1007/s00464-016-4943-5
105. Sgaramella LI, Gurrado A, Pasculli A, et al. The critical view of safety during laparoscopic cholecystectomy: Strasberg Yes or No? An Italian Multicentre study. *Surg Endosc*. Jul 2021;35(7):3698-3708. doi:10.1007/s00464-020-07852-6

106. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. Jan 2009;24(1):74-9. doi:10.1007/s11606-008-0842-3
107. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. *Journal of occupational and organizational psychology*. 1994;67(3):189-205.
108. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. *Nephron Clinical Practice*. 2010;115(2):c94-c99.
109. Mokkink LB, Prinsen C, Patrick DL, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). *User manual*. 2018;78(1)
110. Hahn S, Puffer S, Torgerson DJ, Watson J. Methodological bias in cluster randomised trials. *BMC Med Res Methodol*. Mar 2 2005;5:10. doi:10.1186/1471-2288-5-10
111. Dalmaijer ES, Nord CL, Astle DE. Statistical power for cluster analysis. *arXiv preprint arXiv:200300381*. 2020;
112. Clarke P. When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*. 2008;62(8):752-758.
113. Mulrow CD. Rationale for systematic reviews. *BMJ*. Sep 3 1994;309(6954):597-9. doi:10.1136/bmj.309.6954.597
114. Higgins JP, Thomas J, Chandler J, et al. *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons; 2019.
115. Urlings MJE, Duyx B, Swaen GMH, Bouter LM, Zeegers MP. Citation bias and other determinants of citation in biomedical research: findings from six citation networks. *J Clin Epidemiol*. Apr 2021;132:71-78. doi:10.1016/j.jclinepi.2020.11.019
116. Joober R, Schmitz N, Annable L, Boksa P. Publication bias: what are the challenges and can they be overcome? *J Psychiatry Neurosci*. May 2012;37(3):149-52. doi:10.1503/jpn.120065
117. Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. *CMAJ*. Apr 10 2007;176(8):1091-6. doi:10.1503/cmaj.060410
118. Fagard RH, Staessen JA, Thijs L. Advantages and disadvantages of the meta-analysis approach. *J Hypertens Suppl*. Sep 1996;14(2):S9-12; discussion S13. doi:10.1097/00004872-199609002-00004